AIR FORCE PERSONNEL RESEARCH ISSUES

A Manager's Handbook

September 2007

Report Documentation Page		*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **01 SEP 2007**	2. REPORT TYPE **Final**	3. DATES COVERED **01-10-2006 to 30-09-2007**	
4. TITLE AND SUBTITLE **Air Force Personnel Issues. A Manager's Handbook**		5a. CONTRACT NUMBER	
		5b. GRANT NUMBER	
		5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) **Jacobina Skinner; Nancy Thompson; Kenneth Schwartz; Johnny Weissmuller**		5d. PROJECT NUMBER	
		5e. TASK NUMBER	
		5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Operational Technologies Corporation,4100 N.W Loop 410, Suite 230,San Antonio,TX,78229-4253**		8. PERFORMING ORGANIZATION REPORT NUMBER **; AFCAPS-FR-2010-0017**	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **Air Force Personnel Center - Strategic Research and Analysis Branch, 550 C St West Ste 45, Randolph AFB, TX, 78150-4747**		10. SPONSOR/MONITOR'S ACRONYM(S) **HQ AFPC/DSYX**	
		11. SPONSOR/MONITOR'S REPORT NUMBER(S) **AFCAPS-FR-2010-0017**	

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Handbook is designed to provide Manager's with convenient access to information about completed Air Force research on military personnel issues. Handbook covers topics and provides brief summaries on reseacrh related to both enlisted and officer personnel systems.

15. SUBJECT TERMS
Officer, Enlisted, Personnel Systems, ASVAB, MAGE, Person-Job Match, Learning Abilities, AFOQT, BAT, TBAS, SDI+, Policy Capturing, Policy Specifying, Bias, Commissioning, Performance appraisal

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**9**	**156**	

INTRODUCTION

This handbook is designed to provide managers with convenient access to information about completed Air Force research on military personnel issues. The handbook consists of brief summaries of research topics addressing enlisted and officer personnel systems. To make the handbook a practical resource for managers, emphasis is given to describing the background and major findings from the research areas. Potential applications to current personnel systems issues and recurring problems are highlighted. Findings from many of the completed research programs are directly germane to today's Air Force environment and current force management.

The handbook is intended to serve several purposes. The handbook is primarily a tool to help the target audience – military personnel managers – gain familiarity with major research areas with ease and efficiency. The handbook is also designed to help the staff at the Air Force Personnel Center, Force Management Liaison Office (HQ AFPC/DPST) in formulating and refining future research agendas. The summaries provide sufficient detail to insure that future research programs are designed to build on and extend past existing efforts and that limited resources are not wasted "reinventing the wheel" by duplicating completed research studies. A final purpose is to support HQ AFPC/DPST in its ongoing initiative to preserve Air Force research, principally that completed by the Air Force Human Resources Laboratory (AFHRL) which was disestablished in 1999. This handbook is one of several ways being used to record the organization's legacy so that it is available to managers and researchers.

Technical reports, bibliographies, journal articles, papers in conference proceedings, internet references, and books were reviewed to develop a list of topics for the handbook. Comprehensive coverage of important personnel research areas was the goal. An historical perspective was taken, and research from early eras is described to provide a context for understanding how research programs and Air Force personnel systems evolved. Readers will find brief descriptions of research efforts dating from World War I, as well as studies completed as recently as last year. Most topics are contemporary. The vast majority of military personnel research has been conducted on tests and procedures for screening applicants for service and assigning qualified candidates to Air Force specialties. This handbook reflects the emphasis on personnel selection and classification systems. Summaries of research conducted on other phases of the personnel life cycle like promotion, attrition, force utilization, and retention are also included. The handbook is not exhaustive, and there are several notable areas in which AFHRL conducted large research programs and for which summaries are not provided in this handbook. One area is training research. Another is occupational measurement. Both were beyond the scope of this project. In the case of occupational measurement, the job survey and analysis technologies developed by AFHRL were transitioned to operational programs in 1970 and are accomplished today by the Occupational Measurement Squadron (OMS).

The handbook is organized in three sections: (1) enlisted personnel systems, (2) officer personnel systems, and (3) research methodologies. Managers can find

summaries of specific topics in these sections by consulting the Table of Contents or by using the alphabetical keyword index at the end of the handbook.

A separate section was devoted to research methodologies that are uniquely applicable to military personnel research. Military personnel managers understand the complexities and interrelationships among force management components for recruiting, selection, classification, promotion, and reenlistment. Scientists design research methodologies which are specifically tailored to address some of the large-scale and complicated issues inherent in military systems. A few of the most important methodologies are described in the third section. Researchers within the Air Force and in contracting organizations should be familiar with the methodologies. Some of the methods have been adopted by the private sector and have made their way into statistical computing packages.

In designing the handbook, the needs of managers were foremost, but the handbook was purposefully constructed in a multi-tiered fashion to be useful to current and future researchers and scientists as well. Three tiers distinguished by level of detail and breath of coverage of a research topic are offered. The first tier is the brief summary prepared as a high level overview for managers. The summaries are one to two pages in length and conclude with a reference to one or more supplemental readings. The supplemental readings represent a second tier in terms of the depth of coverage, and they are an important resource for managers who are interested in learning more about a topic. Several of the second-tier documents are papers written specifically for this project in the past year, and they cover research topics in greater detail. The papers vary in length depending on the topic; some are as short as three pages and others are as long as 35 pages. Other types of documents, which are designated as supplemental readings, are technical reports or journal articles, and for a few topics, book chapters. The third and final tier is provided by the citations in the reference lists of the suggested readings. These reference lists point to scores of individual studies completed by the Air Force, other Services, contractors, and academicians. For the most part, the third-tier reference lists will be of interest primarily to scientists who often require detailed information about research methods and results from individual studies.

Materials for the first and second tiers are part of the handbook. The first tier is represented by the brief summaries of topics. To provide managers with easy and quick access to the second tier documents, electronic copies of most supplemental readings were placed on a compact disk. The disk is inserted in a pocket at the back of the handbook. Each reference in the handbook ends with an alphanumeric code identifying the corresponding electronic file on the disk. The code used for enlisted topic references begins with E, officer topics with O, and research methods with RM.

Users of the handbook who are interested in locating individual studies, the third-tier documents, should use the citations in the reference lists as a starting point. Many of the technical reports published by the Air Force Human Resources Laboratory are available from HQ AFPC/DPST at Randolph Air Force Base. As part of their effort to preserve the history of the Air Force research program, more than 1,000 technical reports

have been scanned into electronic files which are available to qualified requesters. The technical reports can also be ordered or in the case of many recently published reports, downloaded from online sites maintained by the Defense Technical Information Center (DTIC) or National Technical Information Service (NTIS). Other possible sources are the Educational Resources Information Center (ERIC) and online subscription services for refereed journals. Using a search engine, a source for many of the documents, including books and book chapters, can be readily determined. Besides online resources, the documents can be found in academic libraries maintained by the Air Force or by the private sector, usually colleges or universities.

The majority of research summarized in this handbook was completed by the Air Force Human Resources Laboratory and by its predecessor and successor organizations. Studies by other organizations including the Air Force Institute of Technology, Department of Defense, personnel research functions in the other Services, and by government contractors are incorporated as well.

In the handbook, we consistently invoke the name "Air Force Human Resources Laboratory," although officially the designation was used only from 1968 to 1991. However, it was the name used for the longest period of time and is the one that has the greatest familiarity to professionals, in and out of the government, with an interest in military psychology. The antecedents of AFHRL can be traced to the Psychological Research Units of the Aviation Psychology Program in the Army Air Corps during World War II. After the Air Force became a separate service in 1947, AFHRL was called Human Resources Research Center (1949-1953), Personnel and Training Center (1954-1958), Personnel Laboratory (1958-1962), and Personnel Research Laboratory (1962-1968). In 1991, the name Air Force Human Resources Laboratory was "retired", and the mission was absorbed by successor organizational units within the Armstrong Laboratory (1991-1996) and the Air Force Research Laboratory (1997-1999). Users of the handbook will find citations for studies published by scientists assigned to all the named organizations.

In 1999, the personnel research function in the Air Force was eliminated, and no organizational entity in the Air Force today has responsibility for research in the domains of personnel selection and classification. Work that continues is conducted primarily under contract, including that sponsored by HQ AFPC/DPST, as well as by small studies and analysis groups within the Air Force. Managers who are interested in ongoing research projects or updates to the topics covered in this handbook are referred to HQ AFPC/DPST.

PREFACE

HQ AFPC/DPST is responsible for operational management of the military testing program for officer and enlisted personnel and sponsored a one-year contract in 2006 to Operational Technologies Corporation, which included the current effort to summarize findings of major personnel research studies conducted by the Air Force. Operational Technologies Corporation appreciates the support of Mr. Kenneth Schwartz, Chief, Force Management Liaison Office, who oversaw the project.

This handbook was prepared by former scientists and research program managers at the Air Force Human Resources Laboratory. The handbook was compiled and edited by Dr. Jacobina Skinner and Ms. Nancy Thompson. Contributors were Dr. William E. Alley, Dr. R. Bruce Gould, Dr. Patrick C. Kyllonen, Dr. Manuel Piña, Jr., Dr. C. Wayne Shore, Dr. Jacobina Skinner, Dr. Mark S. Teachout, Ms. Nancy Thompson, and Dr. Bobby R. Treat.

Preparation of the research summaries and papers was greatly facilitated by the accomplishments of Mr. Johnny Weissmuller, Deputy, Force Management Liaison Office (HQ AFPC/DPST), also a former scientist on the AFHRL staff, and his dedication to preserving the history of the Air Force personnel research program. He made available to the project team electronic copies of over a thousand laboratory technical reports and technical papers, as well as bibliographies and conference papers. These materials were essential in preparing this handbook of research program summaries. Mr. Weissmuller's help is deeply appreciated. We also acknowledge with gratitude the assistance of several individuals who gathered invaluable documents and information for us on a variety of subjects. They include Mr. Kenneth Schwartz, Air Force Personnel Center, Dr. Paul DiTullio, Recruiting Service, Mr. Randy Agee, formerly of the Air Force Occupational Measurement Squadron and now on staff at Operational Technologies Corporation, Dr. John Welsh, Defense Manpower Data Center, Dr. David Alderton, Navy Personnel Research, Studies, and Technologies, Dr. Suzanne Lipscomb, Human Systems Center, and Dr. Richard Roberts, Educational Testing Service.

TABLE OF CONTENTS

I. RESEARCH ON AIR FORCE ENLISTED PERSONNEL SYSTEMS

Early Enlisted Selection and Classification Tests

Precursors of the ASVAB

Aptitude tests have played an important part in airman selection and classification since the Air Force was established as a separate military service branch in 1947. The development of these tests can be traced from the early tests of World War I to the Joint Service Armed Services Vocational Aptitude Battery (ASVAB) which is currently used for selection and classification (see table).

Aptitude Test	Date Implemented	Used for Selection	Used for Classification
Army Alpha*	1917		
Army Beta*	1918		
Army General Classification Test (AGCT)*	1940	X	
Airman Classification Battery AC-1A	1948		X
Airman Classification Battery AC-1B	1949		X
Armed Forces Qualification Test (AFQT)	1950	X	
Airman Classification Battery AC-2A	1956		X
Armed Forces Women's Selection Test (AFWST)	1956	X	
Airman Qualifying Examination, Form D (AQE-D)	1958	X	X
Airman Qualifying Examination, Form F (AQE-F)	1960	X	X
Airman Qualifying Examination – 1962 (AQE-62)	1962	X	X
Airman Qualifying Examination – 1964 (AQE-64)	1964	X	X
Airman Qualifying Examination – 1966 (AQE-66)	1966	X	X
Airman Qualifying Examination , Form J (AQE-F)	1971	X	X
Armed Services Vocational Aptitude Battery, Form 3 (ASVAB-3)	1973	X	X
Joint Service ASVAB	1976	X	X

* Army Alpha, Beta, and AGCT were used for placement decisions.

The development of military aptitude tests began in World War I with the Army Alpha and Army Beta tests. The Army Alpha was initiated in 1917 as a multiple-choice test for group administration composed of eight subtests that covered verbal, numerical, information, and the ability to follow directions. It was followed in 1918 with the Army

Beta which was a non-verbal counterpart to the Alpha for use with illiterates and those who could not speak English. These tests were used for placement of recruits into jobs.

In 1940, the Army General Classification Test (AFGT) was developed as a test of general learning ability that could be used to help identify those who could not perform in wartime situations, to select recruits, and to place recruits into jobs.

The services continued to use aptitude testing after the war for selection and classification purposes. The Air Force used a two-stage testing process. They used forms of the AGCT until 1950 for selection, but they also started developing tests to be used uniquely for classification. These classification tests, known as the Airman Classification Battery (ACB), were used to determine which of the hundreds of potential military job specialties would be the best match for each recruit. Composite scores derived from combinations of subtest scores from these aptitude tests were used to determine qualifications for various clusters of job specialties. At first, the composites were derived from empirical study of the job specialty characteristics and the validity of the classification tests in predicting airman performance. The third ACB (AC-2A) was the first classification battery to group job specialties into aptitude clusters using mathematical analyses instead of the judgments of job analysts.

Ten classification batteries were used from 1948 until the ASVAB. The first groups of tests were known as the Airman Classification Batteries (ACBs) and the second were the Airman Qualification Examinations (AQEs). The subtests on the first ACB in 1948 were made up of a variety of content areas that produced eight composites (Mechanical, Clerical, Equipment Operator, Radio Operator, Technician Specialty, Services, Craftsman, and Instructor). As the Air Force requirements changed, the subtests and composites changed. With the adoption of the AQE, the number of composites had decreased to four with primary emphasis on verbal and quantitative skills. These composites were Mechanical, Administrative, General, and Electronics. The Air Force still derives these four composites from the ASVAB for use in classification.

In 1948, the DoD requested a single selection test for all the Services. The Armed Forces Qualification Test (AFQT) was put into operation in 1950 and continued as a Tri-Service selection test until 1973 when the Services were again allowed to use their own tests. The Air Force initially used the AFQT for both men and women, but forms 3 and 4 were weighted more heavily with mechanical information. Forms 3 and 4 were found to discriminate against women, so the Air Force was directed to develop a test for women, the Armed Force Women's Selection Test (AFWST), which became operational in 1956 and was used until 1974.

Administration of multiple tests for the Services from 1973 to 1975 was a burden to the examining stations, so the DoD once again called for a single test for the Services. In 1976, the Joint-Service ASVAB became operational and continues as the only test for Armed Services selection and classification. The ASVAB is a culmination of aptitude development that began in World War I when tests were used for placement, followed by the AGCT that was used for selection and placement, and the AFQT that was used for

selection. Classification research begun in 1947 resulted in the development of composites from aptitude tests that could be used to identify recruits who were best qualified to fill the jobs clustered in the composites.

Thompson, N. (2007). *Enlisted selection and classification tests: Precursors of the ASVAB.* San Antonio, TX: Operational Technologies Corporation.   **E-01**

Weeks, J.L., Mullins, C.J., & Vitola, B.M. (1975). *Airman Classification Batteries from 1948 – 1975: A review and evaluation* (AFHRL-TR-75-78). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.   **E-02**

ASVAB in the High Schools - 1960s

The Armed Services Vocational Aptitude Battery (ASVAB) was initiated in 1976 as a tri-Service test to be used for selection and classification of military personnel, but the first ASVABs were Forms 1 and 2 that had been used in the High School Testing Program in the early 1960s.

The Services recognized that the high schools were a rich source of military recruitment. Prior to 1962, there was no operational testing done in the high schools to determine the potential aptitudes of students for military training. In 1962, a high school testing program was inaugurated by the Air Force Recruiting Service. It was felt that testing would be beneficial to both the Air Force and the schools. The test scores provided valuable information about the characteristics of the high school enlistment pool and also gave high school counselors a tool to use to help the students make military career decisions. The initial Air Force high school test was a form of the Airman Qualifying Examination (AQE) that had been used for selection and classification purposes since 1958. Other Services followed with their own high school aptitude test batteries.

In 1966, the Assistant Secretary of Defense for Manpower and Reserve Affairs requested a determination of the feasibility of using a common aptitude test battery that would serve as an instrument for high school counseling. A working group from all the Services developed the ASVAB using the best parts of the various Service classification tests. As a result of the DoD directive for a single test, the first ASVAB for student testing (Forms 1 and 2) was introduced in 1968.

The Armed Forces Qualification Test (AFQT) was initiated for all Services as a selection test for enlisted military personnel in January 1950 and was used until 1973 when the Services were allowed to use their own tests. From 1973 to 1975, the Air Force and Marines used ASVAB 3, a test based on the ASVAB that had been used in the high school testing program. It replaced AQE-J and the AFQT and was composed of nine subtests arranged in order of increasing difficulty. There were 300 items and it required approximately two hours to administer. (See table.) Form 4 was used as an alternate for Form 3 in case of test compromise. For Air Force use, four composites were derived from the subtests to form the indexes for Mechanical, Administrative, General, and Electronics (MAGE) classification composites.

ASVAB 3 Subtests

Tests	Testing Time	Number of Items	Description
Coding Speed	7	100	Assignment of coded numbers by relating them to specific words.
Word Knowledge	10	25	Identification of correct meaning for a stimulus word.
Arithmetic Reasoning	25	25	Verbal presentation of arithmetic problems with simple calculations.
Tool Knowledge	10	25	Identification of proper use of tools.
Space Perception	15	25	Identification of patterns that correspond to solid figures.
Mechanical Comprehension	15	25	Identification of the uses of various mechanical devices.
Shop Information	10	25	Identification of proper use of tools in a shop environment.
Automotive Information	10	25	Evaluates specific knowledge about automobiles and automobile motors.
Electronics Information	10	25	Application of knowledge of electricity and electronics in practical situations.

Thompson, N.A. (2007). *Enlisted selection and classification tests: Precursors of the ASVAB.* San Antonio, TX: Operational Technologies. **E-01**

Vitola, B.M., & Alley, W.E. (1968). *Development and standardization of Air Force Composites for the Armed Services Vocational Aptitude Battery.* (AFHRL-TR-68-110). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **E-04**

ASVAB for Military Enlistment - 1976

The Air Force Human Resources Laboratory was given the initial responsibility for developing the tri-Service ASVAB. The ASVAB reflected the content of classification batteries from the Army, Navy, and Air Force. On January 1, 1976, all Services started using the ASVAB for selection and classification. The use of a single test reduced the burden of administering a test for each branch of the service at the examining stations and allowed applicants to take only one test before deciding on the branch of service they would join. With the implementation of the ASVAB, all Services used it as one-stage testing for selection and classification.

The first tri-Service ASVAB tests were Forms 5, 6, and 7. Form 5 was used in the High School Testing Program and Forms 6 and 7 were used operationally for Military recruitment and classification. The first plan for Forms 5, 6, and 7 called for 15 subtests including 12 cognitive power tests, two perceptual tests and a lengthy Interest Inventory. Items for the Interest Inventory were to be selected from the Army Classification Inventory, the Navy Vocational Interest Inventory, and the Air Force Vocational Interest Choice Examination. The test would consist of 335 items and the interest inventory of 527 items and take about four hours to administer.

The original test was too long for operational use and was restructured by combining some of the subtests and shortening the Interest Inventory to a Classification Inventory which contained only Army questions. Radio Information was merged with Electronics Information and Biological Science and Physical Science were merged to form General Science. The revised test was reduced to 13 subtests, including the Classification Inventory, with a total of 382 questions that required about two and a half hours to administer. The items in each of the subtests were arranged in ascending level of difficulty. The final content of Forms 5, 6, and 7 is shown in the table.

ASVAB 5, 6 & 7 Content

Tests	Testing Time (In Mins)	Number of Items	Description
Attention to Detail	5	30	Speeded addition, subtraction, multiplication and division problems
Numerical Operations	3	50	Speeded numerical calculations
Word Knowledge	10	30	Meaning of selected words
Arithmetic Reasoning	20	20	Arithmetic word problems
Space Perception	12	20	Three dimensional figures from folded patterns
Mathematics Knowledge	20	20	Application of learned mathematics principles
Electronics Information	15	30	Simple electricity and electronics knowledge
Mechanical Comprehension	15	20	Use of mechanical and physical principles
Automotive Information	10	20	Automotive repair and symptoms of malfunctions
Shop Information	8	20	Shop procedures and tools
General Science	10	20	Physical and biological science
General Information	7	15	Geography, sports, history, and automobiles
Classification Inventory	20	87	Interest inventory items designed for the Army

Jensen, H. E., Massey, I.H., & Valentine, L.D. Jr. (1976). *Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6, and 7)*. (AFHRL-TR-76-87, AD-AD-AO37 522). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **E-15**

Thompson, N.A. (2007). *Enlisted selection and classification tests: Precursors of the ASVAB*. San Antonio, TX: Operational Technologies. **E-01**

Armed Services Vocational Aptitude Battery (ASVAB)

Updates to Test Content (1980-2002)

It is necessary to periodically revise the ASVAB to control test compromise, replace obsolete items, and make improvements based on new validity and psychometric advances. The first updated forms of the ASVAB went into effect in 1980 with Forms 8, 9, and 10. Subsequently, updates were made with Forms 11 through the currently used computerized adaptive ASVAB. Some of the forms were used for Military selection and some were used for the High School Testing Program. Beginning with Forms 8, 9, and 10, the ASVAB was reduced from 13 subtests used in Forms 5, 6, and 7 to ten subtests with an administration time of approximately 2 hours and 24 minutes for 334 items. All tests from Form 8 through Form 22 had the same ten subtests with the same testing times. (See table) All of the subtests were power subtests with the exception of Numerical Operations and Coding Speed which were administered as speeded tests.

ASVAB Content 1980-2002

Tests	Testing Time (In Mins)	Number of Items	Description
General Science (GS)	11	25	Physical, life, and earth science
Arithmetic Reasoning (AR)	36	30	Arithmetic Word Problems
Word Knowledge (WK)	11	35	Meaning of selected words
Paragraph Comprehension (PC)	13	15	Understanding of written words from brief paragraphs
Numerical Operations (NO)	3	50	Speeded numerical calculations
Coding Speed (CS)	7	84	Speeded use of a key that matches words and numbers
Auto and Shop Information (AS)	11	25	Automobile tools and shop terminology and practices
Mathematical Knowledge (MK)	24	25	Application of learned mathematics principles
Mechanical Comprehension (MC)	19	25	Use of mechanical and physical principles
Electronics Information (EI)	9	20	Simple electrical and electronics knowledge

Curran, L.T., & Palmer, P. (1990). *Armed Services Vocational Aptitude Battery (ASVAB): Item, overlength and operational length development of Forms 18 and 19.* (AFHRL-TP-89-74). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **E-08**

Palmer, P., Curran, L., & Haywood, C.S. (1990). *Armed Services Vocational Aptitude Battery (ASVAB) Forms 20, 21, and 22: Item development.* (AFHRL-TP-89-77). Manpower and Personnel Research Division, Air Force Human Resources Laboratory. **E-41**

Prestwood, J.S., & Vale, C.D. (1985). *Armed Services Vocational Aptitude Battery: Development of Forms 11, 12, and 13.* (AFHRL-TR-85-16(I)). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **E-42**

Ree, M.J., Mathews, J.J., Mullins, C.J., & Massey, R.H. (1981). *Calibration of Armed Services Vocational Aptitude Battery Forms 8, 9, and 10.* (AFHRL-TR-81-49). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **E-06**

Welsh, J.R., Androlewicz, T.R., & Curran, L.T. (1990). *Armed Forces Vocational Aptitude Battery (ASVAB): Analysis of differential item functioning on Forms 15, 16, and 17.* (AFHRL-TP-90-62). Brooks AFB, TX: Manpower and Research Division, Air Force Human Resources Laboratory. **E-43**

Current ASVAB Test Administration and Use

The ASVAB is currently administered under three conditions. The most common method of administration for Armed Forces enlistment is the computerized adaptive version of the ASVAB known as the CAT-ASVAB which is used at the Military Entrance Processing Stations (MEPS). A paper-and-pencil version of the ASVAB is given where computerized testing is not available. In addition, the high school version of the ASVAB, Forms 23 and 24, is a paper-and-pencil test given at more than 13,000 high schools and post secondary schools through a cooperative program between the Department of Defense and the Department of Education.

The content of the CAT-ASVAB is shown in the table.[1] It is the same content used in the ASVAB since about 1980 with the exception that in 2002, the speeded tests of Coding Speed and Numerical Operations were deleted and replaced with Assembling Objects.

CAT-ASVAB Content

Test	Description
Word Knowledge (WK)	Ability to select the correct meaning of words presented in context and to identify best synonym for a given word.
Paragraph Comprehension (PC)	Ability to obtain information from written passages.
Arithmetic Reasoning (AR)	Ability to solve arithmetic word problems.
Mathematics Knowledge (MK)	Knowledge of high school mathematics principles.
General Science (GS)	Knowledge of physical and biological sciences.
Electronics Information (EI)	Knowledge of electricity and electronics.
Auto and Shop Information (AS)	Knowledge of automobiles, tools, and shop terminology and practices.
Mechanical Comprehension (MC)	Knowledge of mechanical and physical principles.
Assembling Objects (AO)	Ability to figure out how an object will look when its parts are put together.

The Armed Services Qualification Test (AFQT) composite score, used for military enlisted qualification, is derived from the ASVAB. It is a percentile score based on a 99 point scale with 99 being the highest score. The AFQT score is derived from the Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Mathematics Knowledge subtests. For enlistment qualification purposes, AFQT scores are divided into categories with corresponding percentile score ranges.

---

[1] The content of paper-and-pencil version of the ASVAB used in the High School Testing Program differs from that in CAT-ASVAB. Details can be found in the Counselor's Manual.

AFQT Categories and Percentile Score Ranges

Category	Percentile Range	Percent of Civilian Youth
I	93-99	8
II	65-92	28
IIIA	50-64	15
IIIB	31-49	19
IV	10-30	21
V	1-9	9

Congress has passed a law that no Category V applicants can be accepted for enlistment and only 20% of accessions can be from Category IV. The Category IV accessions must also have a high school diploma (no GED). The Services have different minimum requirements for enlistment, but the Air Force requirements are the highest. An Air Force enlistee must have a minimum AFQT score of 36 and have a high school diploma or at least 15 hours of high school credit. AFQT cutoff scores are higher for candidates who do not have a high school diploma or at least 15 hours of high school credit. If a candidate for enlistment has a GED, the candidate must also have a minimum AFQT score of 65. One commentator said that a person is more likely to get struck by lightening than be admitted into the Air Force with a GED.

The AFQT score is not used to determine what kind of jobs the recruit is qualified for. Military job qualification is based on Composite Scores taken from the ASVAB subtests that are unique for each branch of the Service. The Air Force uses four Composites called the MAGE. As shown in the Table, the composite structure was revised in 1998. (see Table). Since Numerical Operations and Coding Speed subtests were replaced with the Assembling Objects subtest in 2002, the Services have been re-evaluating the structure of their classification composites. The Air Force has a study underway that may result in changes to the MAGE structure.

Structure of the Air Force Composites

Air Force Composite	Since 1998	Prior to 1998
Mechanical (M)	AR + MC + AS + 2*VE	MC + GS + 2*AS
Administrative (A)	MK + VE	NO + CS + (WK + PC)
General (G)	AR + VE	AR + (WK + PC)
Electronic (E)	GS + AR + MK + EI	GS + AR + MK + EI

Note: VE (Verbal Expression) = WK + PC

ASVAB Career Exploration Program Counselor Manual (2005, November). North Chicago, IL: HQ USMEPCOM. **E-44**

Earles, J.A., & Ree, M.J. (1998). *Development and evaluation of alternative Air Force Administrative composites from the ASVAB.* Unpublished manuscript. Brooks AFB, TX: Human Effectiveness Directorate, Air Force Research Laboratory. **E-45**

Ree, M. J., & Earles, J.A. (1998). *Development and evaluation of alternative Air Force Mechanical composites from the ASVAB.* Unpublished manuscript. Brooks AFB, TX: Human Effectiveness Directorate, Air Force Research Laboratory. **E-46**

Sellman, W.S. (2004). *Predicting readiness for military service: How enlistment standards are established.* Prepared for The National Assessment Governing Board. **E-47**

ASVAB Norms

The ASVAB is the most widely used multiple aptitude battery in the world. Aptitude tests like the ASVAB must be standardized on a sample of the population that is similar to the individuals who will be taking the test. When the performance of American youth changes significantly, it becomes necessary to update the ASVAB norms to reflect the characteristics of the current youth population. New military aptitude tests also are mathematically calibrated or equated to the older tests to be able to evaluate the distribution of scores on a year-to-year basis in a common metric and provide a consistent explanation for cutoff scores for selection and classification tests. Using equating procedures, the scores in a certain percentile on a new aptitude test theoretically should be equal to the same percentile on the old test.

1944 World War II Mobilization Population Norms

In the case of the early tri-Service ASVAB forms, the tests were normed against the 1944 World War II mobilization population. This was done by administering the new ASVAB forms and an anchor test that had already be normed against the mobilization group. The first forms (5, 6, and 7) of the tri-Service ASVAB were normed using a nationally representative sample of people at the basic training centers and at the Armed Services Entrance and Examination Stations (AFEEs). Examinees took the ASVAB along with the Air Force Qualification Test (AFQT) composite from either the Airman Classification Battery or the ASVAB-3, a test that had been used by the Air Force and Marines prior to the implementation of the tri-Service ASVAB. Form 5 also was administered to over 35,000 male and female students in grades 9-12 who were selected as representative of the national high school population. Form 5 was used for high school administration.

ASVAB Forms 8, 9, and 10, also were normed against the 1944 World War II mobilization population. In this case, the new forms and the Armed Forces Qualification Test (AFQT) Form 7a were administered in a counterbalanced order to 22,400 applicants for military enlistment at geographically dispersed AFEEs.

1980 American Youth Population Norms

In 1980, the Department of Defense and the Military Services along with the Department of Labor sponsored a large-scale project to measure the vocational aptitudes of American Youth. The project was called the Profile of American Youth. The ASVAB Form 8a was administered to about 12,000 men and women who were participants in the National Longitudinal Study (NLS) of Youth Labor Force Behavior. The men and women in the sample were born between January 1, 1957 and December 31, 1964. It was the first time that the ASVAB had been administered to a nationally representative sample and the data base was designed to be projected to represent the entire population born in these seven years. Clearly, some characteristics of the youth population had changed over the 36 years since the 1944 mobilization norms were established. In addition, the mobilization norms were based on data collected from males. For more information, see the summary on *The Profile of American Youth – 1980* in this Handbook.

The data from the 1980 survey became the basis for norms for the ASVAB beginning on October 1, 1984 with Forms 10, 11, and 12. The new forms were administered along with Form 8a which was the form given in the 1980 survey. A total of 14, 971 examinees were tested at the Recruit Training Centers and 78,182 examinees were tested at the Military Entrance Processing Stations. Data gathered from these administrations were used to equate Forms 10, 11, and 12 to Form 8a. ASVAB forms were normed against the 1980 population until 2004.

1997 Profile of American Youth Norms

In July, 2004, the Services implemented new norms for the ASVAB, replacing the 1980 Profile of American Youth norms. These current norms were based on the 1997 Profile of American Youth survey conducted by the Department of Defense and the Department of Labor. The computerized adaptive test version of the ASVAB (CAT-ASVAB) was administered to a nationally representative sample of youth 18-23 years old and a sample of 10th, 11th, and 12th grade students. ASVAB tests are now equated back to the CAT-ASVAB that was administered in the 1997 Profile of American Youth survey. This survey found that the 1997 youth scored higher on verbal and math areas and lower on technical areas than the 1980 youth.

Jenson, H.E., Massey, I.H., & Valentine, L.D. (1976). *Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6, and 7)* (AFHRL-TR-76-87). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **E-15**

Martin, C.J. & Welsh, J.R. (1999). Comparison of 1980 and 1997 ASVAB norming Procedures. Monterey, CA: *Proceedings of the 41st Annual Conference of the International Military Testing Association*, p. 329. **E-05**

Ree, M.J., Mathews, J.J., Mullins, C.J., & Massey, R.H. (1981). *Calibration of Armed Services Vocational Aptitude Battery Forms 8, 9, and 10* (AFHRL-TR-81-49). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **E-06**

Ree, M.J., Welsh, J.R., Wegner, T.G., & Earles, J.A. (1985). *Armed Services Vocational Aptitude Battery: Equating and implementation of Forms 11, 12, and 13 in the 1980 youth population metric* (AFHRL-TP-85-21). Brooks AFB, TX: Manpower and Research Division, Air Force Human Resources Laboratory. **E-07**

CAT-ASVAB (Computerized Adaptive Testing (CAT) Version of the ASVAB)

The Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) is now used operationally at all Military Entrance Processing Stations (MEPS). It is a replacement for the paper-and-pencil version of the ASVAB (P&P-ASVAB). The two ability testing methods are based on different theories of individual differences measurement: Item Response Theory for CAT-ASVAB and Classical Test Theory for P&P-ASVAB. With a conventionally administered, printed test, every examinee takes the same items, typically in the same order regardless of the appropriateness of each item for each examinee's ability level. In adaptive testing, the test is tailored to the ability level of each examinee as information on item responses is gathered dynamically during actual test administration. At the beginning of an adaptive test, an item of average difficulty is given because the test taker is assumed to be of average ability. If the examinee responds correctly, a more difficult item is presented next. This process continues until the examinee does not respond correctly. Then, an item is chosen of a difficulty level that falls between that of the last item answered correctly and the item answered incorrectly. In this way, the adaptive testing software continuously selects items, scores the responses, updates estimates of the examinee's ability level, and identifies the next best item for administration to that particular test taker.

The CAT procedure offers several advantages. Test administration time is reduced because through the adaptive testing process, an accurate estimate of an examinee's ability is obtainable with fewer test questions than are required with the P&P-ASVAB. The CAT-ASVAB is less susceptible to compromise and coaching. Sharing of item content among applicants and recruiters is less "profitable," because, in essence, each applicant receives his/her own individualized test containing test items that are uniquely tailored to his/her ability level. Scoring errors (from hand or scanner scoring of P&P-ASVAB) are reduced. Test security is improved; there are no test booklets to be stolen or marked. Further, computer administration provides a less costly method of trying out experimental items to update the item pool and is done in a way that is transparent to examinees.

Research on the development of the CAT-ASVAB began in 1979. Data from over 400,000 test-takers collected over a 20-year period were used to address a variety of crucial research issues on system design and delivery (hardware, software) and psychometric development and evaluation topics. The project was a joint-service effort overseen by the Office of the Assistant Secretary of Defense for Force Management and Personnel. The Navy personnel research laboratory served as executive agent for the DoD with responsibility for the research and development program. The Army laboratory procured/leased the delivery system, and the Air Force Human Resources Laboratory (AFHRL) developed the CAT-ASVAB item pools. Among the important findings, the research demonstrated that the CAT-ASVAB measures the same constructs and achieves the same level of predictive validity as the P&P-ASVAB.

Today, the Under Secretary of Defense for Personnel and Readiness (USD(P&R)) sets policy on military personnel accession testing. The Defense Data Manpower Center (DMDC) has responsibility for CAT-ASVAB research and development. The Secretary of the Army is the executive agent for test administration at the MEPS.

Sands, W.A., Waters, B.K., & McBride, J.R. (Eds.) *Computerized adaptive testing: From inquiry to operation.* Washington, D.C.: American Psychological Association. (Also published as HumRRO FR-EADD-96-26) **E-09**

Enhanced Computer Administered Tests (ECAT)

Recognizing that the widespread availability of CAT-ASVAB computers would facilitate experimentation with new types of tests that could not be administered via paper-and-pencil, the OSD directed the services to begin the Enhanced Computer Administered Tests (ECAT) project in 1989. The purpose was to identify new content to improve ASVAB validity, resulting in cost savings through improved selection and classification of enlisted personnel, reduced school attrition rates and improved on-the-job performance. The services jointly identified nine tests measuring spatial ability, working memory capacity, psychomotor skills, and perceptual speed to form the ECAT battery. The tests were evaluated in studies with Air Force, Army, and Navy samples for incremental validity to the ASVAB, adverse impact reduction, and reliability. Criterion measures for validation studies emphasized "hands-on performance" measures, whenever possible, in addition to technical school grades traditionally used to validate the ASVAB. The "hands-on performance" measures included information on practical skills taught in shop, laboratory, simulator, or other exercises during training courses. One of the ECAT subtests called Assembling Objects (AO) has been added to the ASVAB. The AO subtest is a spatial construction test that includes semi-mechanical items and items that require mental rotation of objects. When AO was added, two ASVAB subtests – Numerical Operations and Coding Speed – were removed. The Air Force is currently evaluating the content changes and need to update classification composites.

Wolfe, J.H., Alderton, D.L., Larson, G.E., Bloxom, B.M., & Wise, L.L. (1997). Expanding the content of CAT-ASVAB: New tests and their validity. In Sands, W.A., Waters, B.K., & McBride, J.R. (Eds.) *Computerized adaptive testing: From inquiry to operation*. Washington, D.C.: American Psychological Association. (Also published as HumRRO Report No. FR-EADD-96-26.) **E-09**

ASVAB Research Topics

Criterion-Related Validity

The answer to the question of whether the ASVAB is a valid predictor of military performance is an unequivocal "yes." Hundreds of studies show that the AFQT, subtests, Service-specific classifications composites, and various ability factors extracted from the test battery are valid predictors of recruits' training and job performance. The results pertain to both paper-and-pencil and computer-adaptive test formats.

Traditionally, the ASVAB is validated against grades obtained in technical training. Positive relationships between test scores and training achievement levels have been found in a host of military technical schools, for a variety of jobs, and in all the Services.

In a recent Air Force study of the 100 most populated enlisted specialties, the ASVAB subtests were found to predict technical training final course grades. The median multiple correlation (predictive validity) was $R = .45$. The range was .25 to .62 and even the smallest R was highly statistically significant ($p < .0001$). These correlations are very favorable when compared to predicting academic performance by the most popular commercial selection test, the Scholastic Aptitude Test (SAT). On their web site, the Educational Testing Service (ETS) points to an ETS summary of the annual validities of the SAT. The study predicted freshman academic performance (GPA) over each of 15 years and the multiple regression results have a median R of .47 and a range of .41 to .57. The validities obtained for the SAT are comparable to those found for the ASVAB.

The ASVAB also predicts important military criteria outside the schoolhouse. Studies of job performance measures show the ASVAB relates to how well airmen perform technical aspects of their jobs, including hands-on tasks. Further, ASVAB scores are predictive of whether individuals complete their initial enlistment or become premature attritions. Enlistees scoring lower on the ASVAB are more likely to attrit prematurely, thus providing less mission support and less return on the Air Force's recruiting and training investment. Also, data indicate that ASVAB scores are related to the number of productive man hours over the first four years in service for those who complete their first four years. Airmen who score higher on the ASVAB are more productive members of the force.

Periodic checks on ASVAB validity are an integral part of the testing program. Air Force studies are ongoing to update relationships with recent first-term attrition and productivity indices. At the DoD-level, extensive reviews of technical issues affecting validity coefficients and availability of criterion measures have recently been completed in support of validation efforts by the Services.

McCloy, R.A., Campbell, J.A., Knapp, D.J., Strickland, W.J., & DiFazio, A. S. (2006). *A framework for conducting validation research with the Armed Services Vocational Aptitude Battery (ASVAB)*. Alexandria, VA: HumRRO. **E-10**

Welsh, J.R., Jr., Kucinkas, S.K., & Curran, L.T. (1990). *Armed Services Vocational Aptitude Battery (ASVAB): Integrative review of validity studies* (AFHRL‑TR‑90‑22). Brooks AFB, TX: Air Force Human Resources Laboratory. **E-11**

Test Bias

One important standard a personnel selection test must meet is that it be unbiased with respect to minority subgroups. The proper concept of bias is somewhat technical and is not the same as the colloquial use of the term to refer merely to differences in subgroup test performance.

To understand the proper definition of test bias, one must consider the relationship between predictive test scores and later measures of performance. Personnel selection tests are useful to the extent that they predict eventual job performance. This predictive ability of the test allows for the establishment of minimum scores for accessions and for assignments to various jobs.

Bias occurs when a subgroup's actual performance is under-predicted (underestimated) by the personnel selection test. If a subgroup performs better on the criterion than predicted by a selection test, then the use of that selection test is not equitable for that subgroup.

The historical sensitivity of this issue determines that a selection test may not be used if its use results in inequitable treatment for minority subgroups, with such subgroups defined as females and ethnic or racial minorities.

The Air Force and the Army conducted studies to determine possible bias in its tests used to determine whether applicants were qualified to enlist (such as the current Armed Services Vocational Aptitude Battery). Two large-scale Air Force studies (Guinn, Tupes, & Alley (1970) and Shore & Marion (1972)) showed no bias against blacks. Similar studies conducted by the Army (Maier & Fuchs, 1973) and by Joint Service testing researchers (Wise et al., 1992) also reported no practical levels of bias against racial and gender minorities.

Based on the best available evidence, there is no reason to believe that there is any bias disfavoring minorities by selection tests used in the Air Force. However, newly developed tests need to be reviewed to determine that they don't under-predict the eventual performance of minority subgroups.

Guinn, N., Tupes, E.C., & Alley, W.E. (1970). *Cultural subgroup differences in the relationships between Air Force aptitude composites and training criteria* (AFHRL-TR-70-35). Brooks AFB, TX: Air Force Human Resources Laboratory. **E-40**

Maier, M.H., & Fuchs, E.F. (1973). *Effectiveness of selection and classification testing* (ARI-RR-1179, AD0768168). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences. **E-49**

Shore, C.W., & Marion, R. (1972). *Suitability of using common selection test standards for Negro and White airmen* (AFHRL-TR-72-53). Lackland AFB, Texas: Personnel Research Division, Air Force Human Resources Laboratory. **E-50**

Wise, L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L, & Divgi, D.R. (1992). *Sensitivity and fairness of the Armed Services Vocational Aptitude Battery (ASVAB) technical composites*. Seaside, CA: Defense Manpower Data Center. **E-51**

Research on Low-Aptitude Recruits: Project 100,000

The DoD has on several occasions admitted large number of low-aptitude individuals into the military services. Every national mobilization of manpower has produced the need to relieve the pressure on the recruiting pool by more extensive utilization of low-aptitude personnel. This occurred during World War II and the Korean conflict. In response to the escalating manpower needs brought about by the Vietnam War, another program was initiated. Called Project 100,000, it was led by Secretary of Defense Robert S. McNamara and was tied to President Lyndon B. Johnson's War on Poverty. The program received its name from the goal of accepting 100,000 men per year who did not meet mental standards. Between 1966 and 1971 about 354,000 "New Standards" accessions were accepted under reduced mental qualifications, many of whom were men in Category IV who scored between the 10th and 30th percentile on the Armed Forced Qualification Test (AFQT). Quotas were established which resulted in 67% of the low-aptitude personnel being assigned to the Army and the remainder being distributed to the other services. Project 100,000 was seen principally as a social program and was very unpopular with military managers.

A major benefit expected from the program by policy-makers was that the remediation and intensive training associated with entry into the service would better enable the "New Standards" personnel to adapt to both the military environment and future civilian life. The program prompted numerous studies on the performance, trainability, and utilization of low aptitude personnel and comparisons with populations of men who either met mental standards or who were non-veterans who did not serve in the military.

In 1976, the Air Force Human Resources Laboratory reviewed 62 separate studies conducted by the Services or by DoD on "New Standards" personnel and completed between 1966 and 1975. The major finding was that although the "New Standards" personnel did not perform as well as the more highly educated, more literate, and higher aptitude men in comparison groups, most became highly satisfactory servicemen. They did comparatively well in basic training and occupational training, as well as in terms of promotions and reenlistments but not as well as servicemen who met entry standards. Overall, they required longer to complete training and achieve journeymen status. Generally, the "New Standards" personnel had positive feelings toward their military experience.

Studies on post-service adjustments produced conflicting results. Initial findings from studies completed in the 1970s were that after spending two years in the military, the "New Standards" personnel returned to civilian life and had higher aspirations for education, higher paying jobs, and were in higher skilled occupations than a carefully matched comparison group of non-veterans. However, when the "New Standards" personnel were contacted in a 1986-87 follow-up study by DoD, they were found to be faring less well than their counterparts who had never served in the military. The Project 100,000 participants were more likely to be unemployed, had a lower average level of education, lower income, and a higher divorce rate. Nevertheless, the "New Standards" personnel continued to report positive feelings about their military experience.

The conclusion briefed to Congress by DoD, after completion of studies in the mid-1980s, was that military service does not offer a "leg up" to low aptitude and disadvantaged youth as they seek to overcome cognitive and skill deficits and compete successfully in later civilian life. Throughout Project 100,000, the military services made it clear that they do not regard their role as that of social welfare agency, social equalizer, or as an appropriate avenue for remedying the literacy or skill deficits of America's underprivileged.

Laurence, J.H., Ramsberger, P.F., & Gribben, M.A. (1989). *Effects of military experience on the post-service lives of low-aptitude recruits: Project 100,000 and the ASVAB misnorming* (HumRRO FR-PRD-89-29, Educational Resources Information Center (ERIC) Document Number ED 366 751). Alexandria, VA: Human Resources Research Organization. **E-12**

Ratliff, F.R., & Earles, J.A. (1976). *Research on the management, training, and utilization of low-aptitude personnel* (AFHRL-TR-76-69). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **E-13**

Thompson, N. (2007). *Tracking trends in the enlisted force*. San Antonio, TX: Operational Technologies Corporation. **E-14**

More Research on Low-Aptitude Recruits: Misnorming of the ASVAB

The error causing the ASVAB misnorming was made at the Air Force Human Resources Laboratory (AFHRL), which at the time was lead agency for development of the joint service enlistment test. The normative population identified for converting raw scores to percentile scores was flawed for ASVAB Forms 5, 6, and 7. The result was that, when the forms were put into use in January 1976 for DoD enlistment qualification, recruits were given inflated scores. By the time the error was corrected in October 1980, over 300,000 recruits had been admitted in the military services who would not have qualified for enlistment if the test had been calibrated correctly. The error was discovered by manpower analysts at the Pentagon and by testing specialists in the 1979-1980 timeframe. However, there were anecdotal accounts that complaints from field commanders about a quality decline had begun to surface much earlier.

The impact of the misnorming was not the same for each Service. The enlistment standards differed by Service and by high school graduation status. The misnorming error particularly inflated scores in the lower ability ranges where some Service standards were set. Many recruits thought to be of average aptitude were, in fact, below average or in the Category IV range (AFQT percentiles 10-30). The number of recruits who were erroneously admitted varied by Service; the percentage was highest for the Army (66%), followed by the Navy (17%), and lowest for the Air Force (4%).

Several remedial steps and initiatives followed. The Army decided not to renew enlistment contracts of low-scoring members who entered during the ASVAB misnorming. The AFHRL prepared the revised ASVAB forms with accurate conversion tables which were implemented in October 1980. In addition, Air Force research studies which had used the misnormed scores in analyses were recalled and re-accomplished. The DoD established an advisory panel of testing experts from across the country to conduct an annual review of the ASVAB program. Congress directed that the DoD undertake validation studies to demonstrate that scores on the ASVAB related to how well enlisted personnel performed their jobs. The DoD sponsored a large number of studies of the low-aptitude recruits erroneously admitted to service. The findings were that low-aptitude recruits had higher premature attrition, lower retention rates, and after returning to civilian life, acquired less formal education, had higher divorce rates, and were less satisfied with their jobs compared to non-veterans. They did not differ from non-veterans in terms of employment status, occupational category, or average income.

Jensen, H.E., Massey, I.H., & Valentine, L.D., Jr. (1976) *Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6, and 7)* (AFHRL-TR-76-87). Brooks AFB, TX: Air Force Human Resources Laboratory. **E-15**

Laurence, J.H., Ramsberger, P.F., & Gribben, M.A. (1989). *Effects of military experience on the post-service lives of low-aptitude recruits: Project 100,000 and the ASVAB misnorming* (HumRRO FR-PRD-89-29, ERIC Document Number ED 366 751). Alexandria, VA: Human Resources Research Organization. **E-12**

The Profile of American Youth - 1980

In 1980, the Department of Defense and the Military Services along with the Department of Labor sponsored a large-scale project to measure the vocational aptitudes of American youth. The project was called the Profile of American Youth. The ASVAB Form 8a, which was developed by the Air Force Human Resources Laboratory, was administered to about 12,000 men and women who were participants in the National Longitudinal Study (NLS) of Youth Labor Force Behavior. The men and women in the sample were born between January 1, 1957 and December 31, 1964. It was the first time that the ASVAB had been administered to a nationally representative sample and the data base was designed to be projected to represent the entire population born in these seven years. Assessment of the test scores within and across Services also provided the opportunity to measure enlistees based on a national measure of vocational test performance.

Some of the demographic findings from the Profile of American Youth are summarized below:

1. Average AFQT scores and estimates of reading grade level increased with age.

2. The average AFQT scores for males and females were similar. Males scored higher on Mechanical, Electronics, and General Composites and females scored higher on the Administrative Composite.

3. AFQT scores were higher on the average for White youth than for Black or Hispanic and scores for Hispanic youth were higher than scores for Black youth. The relationships among the races were the same for measures of reading grade level and for the four Air Force classification composites.

4. AFQT test performance was strongly correlated with formal education. Non-high school graduates had the lowest average AFQT scores and graduates had the highest average AFQT scores. Youth holding GEDs scored between these two groups.

5. Scores for youth on the AFQT were found to increase in direct correspondence to the amount of formal education completed by their mothers.

6. Youth from the New England and West North Central regions of the country had the highest average AFQT scores and youth from the southern divisions had the lowest average scores.

Using the enlistment standards for 1981, it was estimated that 62.6% of the total population ages 18-23 years would have qualified for the Air Force. Broken down by race, 71.3% of White youth would have qualified, 21.3% of Black youth would have qualified, and 37.5% of Hispanic youth would have qualified. The percent of youth who would have qualified for the other Services were: 76.3% would have qualified for the

Army, 75% would have qualified for the Navy, and 72.4% would have qualified for the Marines.

The military used the results of the study to assess the attributes and trainability of the military-age population by geographic area and social category, to estimate the effects of modifications to the aptitude and education standards, to track the vocational aptitudes and attitudes toward the military, and to gauge the comparative aptitudes of different demographic subgroups.

Bock, D.R., & Moore, E.G.J. (1984). *The Profile of American Youth: Demographic influences on ASVAB test performance* (DTIC Report No. AD-A125 830). Washington, D.C.: Office of the Assistant Secretary for Defense (Manpower, Installations, and Logistics). **E-48**

Eitelberg, M.J., Laurence, J.H., Waters, B.K., & Perelman, L.S. (1984). *Screening for Service: Aptitude and education criteria for military entry*. Washington, D.C.: Office of Assistant Secretary of Defense (Manpower, Installations and Logistics). Electronic copy not available.

Estimating Reading Ability from the ASVAB

Annually, the Office of the Under Secretary of Defense (Personnel and Readiness) prepares a report to Congress on the demographic characteristics and quality of accessions. Reading ability is one of several quality indicators reported. In 2004, for example, Congress was informed that reading levels were higher in the enlisted military than in the non-military sector. Further, all services showed improvements from 1984 to 2004. The mean reading grade level of Air Force accessions increased from 10.5 (reading grade level at the fifth month of the 10th grade) to 11.4 (fourth month of the 11th grade).

Reading grade levels in the report are estimated from applicants' scores on verbal subtests in the Armed Services Vocational Aptitude Battery (ASVAB). In the 1980s, a large scale equating study for the ASVAB and reading tests was recommended by the Department of Defense (DoD) Joint-Service Selection and Classification Working Group (JSSCWG) and completed under contract by the Human Resources Research Organization (HumRRO). Over 20,000 military applicants were tested at the MEPS on the ASVAB and several reading tests. The purpose was to obtain an ASVAB-anchored estimate of applicants' reading grade level. The Verbal composite, which contains Word Knowledge and Paragraph Comprehension subtests, was selected as the most accurate anchor for the reading ability scale. The ASVAB conversion tables produced in this study are used to make the annual reports to Congress. They have also been used to describe the reading ability of Air Force applicants assigned to different specialties.

Because of the high verbal content in the Armed Forces Qualification Test (AFQT) on which applicants are required to meet minimum scores, most examinees with very low reading skills are screened out and do not enter service. There is no formal minimum entry requirement on the reading ability scale.

Skinner, J. (2007). *Air Force Reading Abilities Test (AFRAT) and related topics.* San Antonio, TX: Operational Technologies Corporation. **E-17**

Air Force Special Purpose Tests

Enlistment Screening Test (EST)

The military services use screening tests to reduce enlistment processing costs. Recruiters administer the screening tests locally, identify applicants who likely will meet service mental qualifications, and arrange for them to travel to central Military Entrance Processing Stations (MEPS) for additional testing. Transportation and boarding costs are avoided for applicants whose probability of meeting entrance standards is extremely low. The traditional use of screening tests by recruiters in all military services has been to predict the likelihood an applicant will meet or exceed the minimum Armed Forces Qualification Test (AFQT) score required for enlistment eligibility. The AFQT is a composite score derived from the Armed Services Vocational Aptitude Battery (ASVAB).

For many years, recruiters relied on the Enlistment Screening Test (EST) for this purpose. The most recent version of a paper-and-pencil EST was developed for the Marine Corps by the Center for Naval Analysis (CNA). The Defense Advisory Committee (DAC) on Military Personnel Testing and the other services expressed interest in expanding CNA's effort to construct a joint-service screening test. After Navy and Air Force data were collected and analyzed, two parallel screening tests called EST A and B were implemented. Each form contained 65 items total across Word Knowledge, Arithmetic Reasoning, and Math Knowledge content areas. The time limit for test completion by applicants was 45 minutes. The joint service EST, along with the expectancy tables for AFQT scores, were printed and distributed to all services in February 1989. These forms are currently authorized for use by Air Force recruiters (AFPT Catalog, 1 June 2006). Air Force recruiters also use a computerized (page turner) EST (Version 1.0). The test is DOS-based and consists of four parts: Word Knowledge (18 items), Arithmetic Reasoning (15 items), Paragraph Comprehension (8 items), and Math Knowledge (13 items). Total administration time is 39 minutes. Recruiters are also authorized to use another computerized screening test with item selection based on adaptive testing techniques; this test is named the Computerized Adaptive Screening Test (CAST).

Skinner, J. (2007). *Enlistment Screening Test (EST) and Computerized Adaptive Screening Test (CAST)*. San Antonio, TX: Operational Technologies Corporation. **E-19**

Computerized Adaptive Screening Test (CAST)

The military services use screening tests to reduce enlistment processing costs. Recruiters administer the screening tests locally, identify applicants who likely will meet service mental qualifications, and arrange for them to travel to central Military Entrance Processing Stations (MEPS) for additional testing. Transportation and boarding costs are avoided for applicants whose probability of meeting entrance standards is extremely low.

The traditional use of screening tests by recruiters in all military services has been to predict the likelihood an applicant will meet or exceed the minimum Armed Forces Qualification Test (AFQT) score required for enlistment eligibility. The AFQT is a composite score derived from the Armed Services Vocational Aptitude Battery (ASVAB).

For many years, recruiters relied on a paper-and-pencil Enlistment Screening Test (EST). In the early 1980s, development began on the Computerized Adaptive Screening Test (CAST). The purpose of CAST was a quicker and easier screening test for recruiters to administer than the paper-and-pencil EST, which required hand-scoring and hand-conversion to an estimated AFQT score. With CAST, computer software is used to tailor test difficulty to the examinee's ability. Adaptive tests typically achieve the measurement precision of conventional, non-adaptive tests with half the number of items. The validation efforts revealed that CAST predicted AFQT at least as accurately as the EST, was more efficient, reduced the administrative burden on recruiters, and was less susceptible to test compromise. The CAST was first implemented in the Army. Later research resulted in revisions to improve its psychometric properties, modifications for use on a succession of microcomputer models, and changes to prepare the test for joint service use. CAST Version 5 is authorized for use by Air Force recruiters. Examinees are tested on Word Knowledge and Arithmetic Reasoning items, and most test-takers are finished in about 25 minutes. Recruiters are also allowed to administer the Enlistment Screening Test (EST) instead of the CAST.

Sands, W.A., Gade, P.A., & Knapp, D.J. (1997). The Computerized Adaptive Screening Test. In Sands, W.A., Waters, B.K., & McBride, J.R. (Eds.) *Computerized adaptive testing: From inquiry to operation*. Washington, D.C.: American Psychological Association. (Also published as HumRRO FR-EADD-96-26) **E-09**

Skinner, J. (2007). *Enlistment Screening Test (EST) and Computerized Adaptive Screening Test (CAST)*. San Antonio, TX: Operational Technologies Corporation. **E-19**

Electronic Data Processing Test (EDPT)

The EDPT is used to classify airmen into Air Force specialties requiring computer operations and programming skills. In recent years, the test has been used for AFSC 3C0X2, Communications – Computer Systems Programming, and Reporting Identifiers 9S100, Scientific Measurement Technician and 9S200, Applied Sciences Technician. To qualify for the specialties, airmen must meet minimum qualifying scores on the EDPT, as well as cognitive ability requirements on the Armed Services Vocational Aptitude Battery (ASVAB). The Marine Corps also uses the EDPT for assignments to jobs in the computer field. Anecdotes found on http://usmilitary.about.com indicate that the EDPT has the reputation among examinees as being one of the most difficult tests administered at the Military Entrance Processing Stations.

The history of the EDPT dates to 1961 when the Strategic Air Command requested an Air Force-developed test for selecting personnel for computer programming training. At the time there were no formal technical training courses in electronic data processing equipment repair or programming. Air Force personnel were trained by customer service representatives of commercial equipment manufacturers. In some instances, before training began, the manufacturer would administer their own selection test.

The EDPT was constructed to resemble and was normed to the International Business Machines (IBM) Programmer Aptitude Test (PAT), which during the 1960s was the most widely used selection test for computer programmers and systems analysts in American and Canadian companies. It consisted of three parts requiring examinees to determine the next number in a series, analogies represented in figures, and solutions to arithmetic problems. Except for the addition of a Verbal Analogies subtest, the ability areas tested in the EDPT were the same as those in the IBM PAT. The four subtests of the EDPT (Arithmetic Reasoning, Figure Analogies, Verbal Analogies, and Number Series) are administered as a 90-minute power test. Each subtest contains 30 items. Although the EDPT developed in the 1960s is still in operational use, a review of the test items in the 1990s revealed that the item content was not obsolete.

There have been few studies on the EDPT but those published indicate that it is a valid predictor of technical training performance. However, the results are conflicting about its value as a classification test and its predictive effectiveness over and above ASVAB measures. An updated study is needed.

Skinner, J. (2007). *Electronic Data Processing Test (EDPT)*. San Antonio, TX: Operational Technologies Corporation. **E-20**

Air Force Reading Abilities Test (AFRAT)

The Air Force Reading Abilities Test (AFRAT) was developed when problems arose with the use of commercial reading tests in the 1970s. Many Air Force organizations had been obtaining commercially published tests, administering them to military personnel, and using the results for assignment to remedial training programs, for aids in counseling students, or for descriptions of reading levels of airmen in various occupational specialties. However, a study on service applicants found differences in the reading grade level (RGL) results from different commercial tests for applicants with the same scores on the Armed Services Vocational Aptitude Battery. The commercial tests were also expensive. Consequently, the Air Force Human Resources Laboratory was tasked with developing a reading ability test with norms appropriate for a military population. The AFRAT was implemented for airmen in 1982, and Air Force agencies were directed to use the forms instead of commercial tests.

The AFRAT consisted of two parallel forms with sections testing vocabulary and comprehension skills. Items were drawn from a large pool of candidate test items written by the Educational Testing Service. The final tests had 45 vocabulary and 40 comprehension items. Testing time was 50 minutes. Studies were conducted which demonstrated the construct validity, reliability, and parallelism of the two forms. Most items were quite easy, a planned test characteristic for detecting reading deficiency. The AFRAT distributions showed negative skew, a desirable feature for identifying low-performing examinees on the reading test. Conversion tables were prepared to place AFRAT scores on a reading grade level (RGL) scale. The RGL scale corresponded to school grade 1 through grade 12 and referenced reading ability, as measured by the AFRAT, to the average ability of students at each school grade/month. For example, an airman with an AFRAT RGL of 9.3 was reading at a level of an average student in the third month of the 9th grade. Separate tables converted AFRAT total score, vocabulary subtest score, and comprehension subtest score to RGL equivalents. Further analyses showed the AFRAT was a valid predictor of airman grades in technical training courses.

The Air Force Reading Abilities Test (AFRAT) is in the active inventory of Air Force personnel tests. Requests for administering the test are processed on a case-by-case basis for identifying enlisted personnel with marginal or inadequate reading ability.

Skinner, J. (2007). A*ir Force Reading Abilities Test (AFRAT) and related topics.* San Antonio, TX: Operational Technologies Corporation. **E-17**

Strength Aptitude Test

The Strength Aptitude Test is used by the Air Force as a classification tool to insure recruits have the physical strength to perform the physical demands of military jobs. Since the 1970s, when the military services began to accept increasing numbers of women, the Air Force has been the only military service branch to use a strength screening tool on a continuing basis as part of recruits' induction procedures at the Military Entrance Processing Stations (MEPS). The Air Force began developing physical strength standards in the early 1970s. The research program originated at what was then the Air Force Aerospace Medical Research Laboratory (AFAMRL) at Wright-Patterson AFB. The program has been continued by the Human Effectiveness Directorate, Air Force Research Laboratory (AFRL/HE), the current OPR for the Strength Aptitude Test. By 1976, a three-level "Factor X" weight lift test was in use. During the next decade, it was updated to a nine-level test which was implemented as the Strength Aptitude Test in 1987.

The Strength Aptitude Test is a weight lifting test performed on an incremental lifting machine similar to the equipment found in fitness centers. The test requires recruits to lift weights starting at 40 pounds. The weight is then increased in 10-pound increments until the recruit (1) cannot complete a lift, (2) asks to stop, or (3) lifts 100 pounds, the maximum requirement of any Air Force job. Job qualifications standards on the Strength Aptitude Test have been established for all Air Force specialties. Originally, the standards were developed by computing an average physical demand weighted by frequency of performance and percent of the AFS members performing a task. Air Force specialties were surveyed for development of strength standards between 1978 and 1982. HQ AFPC/DPPAPC, the OPR for strength classification standards, reported in November 2006 that AFRL/HE resurveys between three and eight AFSs annually to insure job standards are current. Classification standards are set in 10-point increments and range from a low of 40 pounds to a maximum of 110 pounds. About half of current AFSs have a classification standard of a 40-pound lift. As of 2006, the maximum operational standard for any AFS was 100 pounds. Standards are gender-neutral (the same for men and women). There are several sources of data on recruit capabilities to perform the Strength Aptitude Test. Almost all females can lift 40 pounds, the minimum requirement for Air Force jobs. The average weight lifting capability for males is about 114 pounds and for females about 57 pounds.

In 2007, discussions were ongoing about an initiative by the AFRS/CC to eliminate strength aptitude requirements. Potential risks are increased injuries to airmen assigned to specialties where job physical demands exceed their physical capabilities, inadequate performance in job assignments, inequitably higher workloads for men in high demand specialties, and increased burden on supervisors to develop workaround solutions to insure work is accomplished.

Skinner, J. (2006). *Strength Aptitude Test*. San Antonio, TX: Operational Technologies Corporation. **E-22**

Enlisted Classification

Origins of MAGE

The Armed Services Vocational Aptitude Battery (ASVAB) is used for both selection and classification of airmen. As a classification test, it produces four composite scores that are used for assigning airmen to job specialties. These four composites or aptitude indexes are called the MAGE and they cover the job specialties that fall into the areas of performance that are Mechanical, Administrative, General, and Electronics.

The concept of composite scores was developed from the Airman Classification Battery (ACB) that was initiated in 1948 and later the Airman Qualifying Examination initiated in 1958. Composite scores are derived from combinations of scores from the classification battery and are used to determine an airman's qualifications for various job specialties. The early aptitude batteries differed in number and types of abilities measured and the configuration of job clusters. The first form of the ACB yielded eight composites: Mechanical, Clerical, Equipment Operator, Radio Operator, Technical Specialty, Services, Craftsman, and Instructor. The challenges of identifying composites were to predict success accurately within each job cluster and to differentiate those who were likely to be successful in one cluster from those who were likely to be successful in another cluster. The composites had to be valid and they had to be differentially valid. If each of the separately developed composites rank ordered people in the same way across job categories, the core requirement for effective classification would be impossible. As the tests were updated and refined, more emphasis was placed on measuring verbal and quantitative abilities and the number of composites decreased. By the time the AQE-D was introduced as a classification test in 1958, the administration time for the test battery had been shortened and the composites had been reduced to the four MAGE composites.

Early composites were defined by using expert judgments on the properties of job specialties, on the number of composites that would be needed to cover all the job specialties, and on which job specialties would belong to each composite. Researchers did rely on statistical methods including factor analysis and tests of correlation coefficients which, in the earliest studies, were computed by hand. The second ACB, the AC-1B, was the first battery to group specialties into aptitude clusters using mathematical analyses instead of job analysts.

Over the years, advances were made at the Air Force Human Resources Laboratory in analytical capabilities. A technique called hierarchical grouping provided a sophisticated new approach for job clustering. The job clusters produced by using the hierarchical grouping technique closely approximated the pattern of job clusters that had been traditionally defined as Mechanical, Administrative, General, and Electronics job groups and affirmed the procedure of using composite scores for airman classification. However, as Air Force jobs and requirements change and test content is modified, it is necessary to reevaluate the composites to determine the efficacy of each composite and the correct configuration of job specialties for each composite.

Thompson, N.A. (2007). *Enlisted selection and classification tests: Precursors of the ASVAB.* San Antonio, TX: Operational Technologies Corporation. **E-01**

Measuring Occupational Learning Difficulty
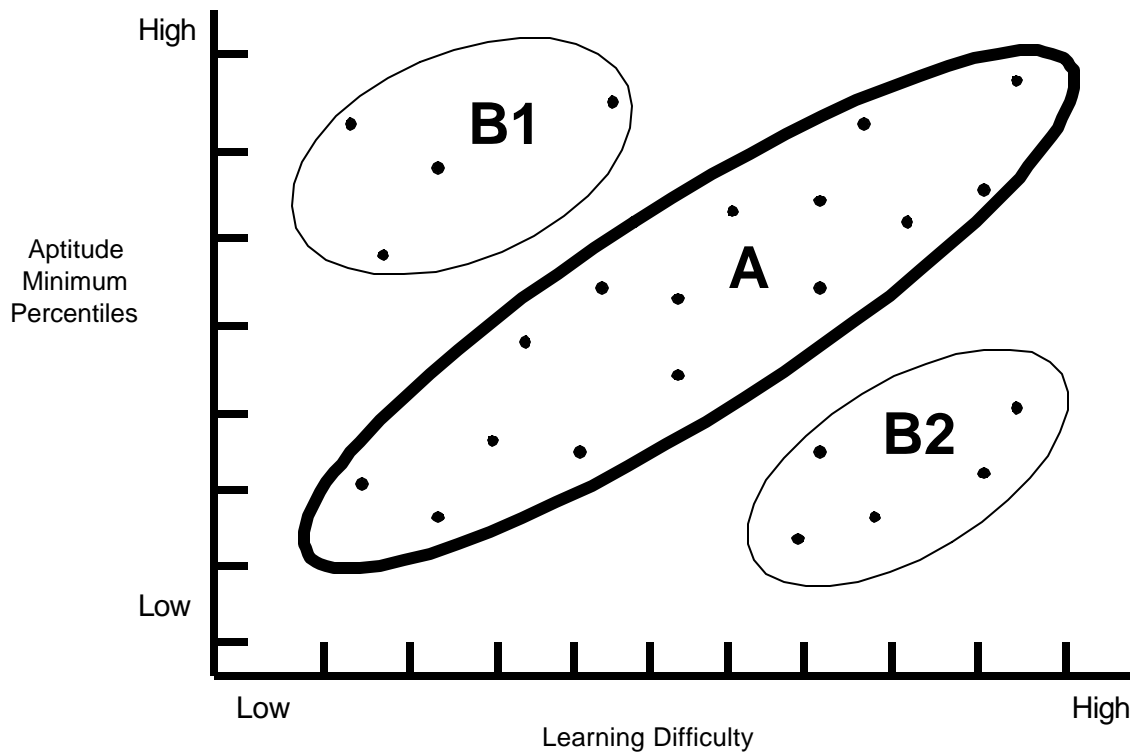## to Establish Aptitude Requirement Minimums

The term "occupational learning difficulty" is associated with a major project completed in the 1980s by AFHRL. The purpose was to provide an empirical basis for establishing aptitude requirements for enlisted AFSs. Historically, aptitude minimums were determined primarily based on training outcomes. Relationships between the Armed Services Vocational Aptitude Battery and academic performance and pass/fail rates in training were determined. Aptitude minimums were raised and lowered based on attrition rates and recruiting needs. No systematic decision rules existed for setting the entry standards. In the 1970s concerns were raised about the probable misalignment of aptitude requirements and job demands and the impact on allocation of enlisted talent.

The AFHRL's approach to the problem was to measure occupational mental demand. It was widely recognized that there was tremendous variance in both job demand levels of AFSs and in the learning rates of individual airmen. Beginning in the 1960s, AFHRL research provided procedures for defining and measuring characteristics of tasks and jobs in the Air Force that would correspond to the types of measures obtained on recruits. Among these measures was task learning difficulty which was parallel in concept to the measures of aptitudes for recruits. Task learning difficulty was defined as the time it takes to learn to perform a task satisfactorily. Early research demonstrated that Air Force supervisors could provide highly reliable ratings of relative task difficulty for their career fields.

In order to obtain data for comparing aptitude requirements across Air Force specialties, a technique was designed which allowed the task learning difficulty ratings, formerly available only within specialties, to be calibrated across specialties. The method made use of benchmark scales which allowed direct comparisons between specialties in terms of occupational learning demand. A measure of occupational learning difficulty was obtained separately for over 200 enlisted AFSs, each measure representing how long it takes to learn to perform the occupation satisfactorily. The measures provided a frame of reference for inferring aptitude requirements for occupations in the same aptitude area (Mechanical, Administrative/General, and Electronic).

The value of the occupational learning difficulty measures for establishing the order of aptitude requirement minimums is illustrated by the figure. Each point on the chart represents an AFS and shows the intersection of its learning difficulty (horizontal axis) and aptitude percentile requirement (vertical axis). Specialties in the cluster labeled A are aligned properly with lower demand AFSs having lower aptitude minimums, and higher demand AFSs having higher aptitude minimums. However, the AFSs in clusters B1 and B2 have minimum aptitude requirements that are inconsistent with the difficulty of learning to perform satisfactorily in the occupational. AFSs in cluster B1 have low learning difficulty and high aptitude minimums. To correct the misalignment, standards need to be lowered for AFSs in cluster B1. The opposite occurs for AFSs in cluster B2. These AFSs have high learning difficulty and low aptitude minimums. Aptitude

minimums need to be raised in these AFSs. The amount of adjustment would be the number of percentile points required to shift the AFSs in cluster B1 and B2 to their proper position in cluster A.

In February 1981, a working group was formed to evaluate aptitude requirement minimums for all enlisted specialties. For AFSs with misaligned occupational learning difficulty measures and aptitude standards, the decision was to adjust minimums incrementally by plus/minus 5 percentile points each year until aptitude requirement goals were reached for each job specialty. The working group also considered training and recruiting issues. Revisions were published in the classification regulation in April 1982. Eventually, aptitude requirements were modified for over 100 enlisted specialties.

Although several research studies addressed the need for alternate and more efficient methods of measuring occupational learning difficulty, little headway was made in designing a replacement methodology. Because of the importance of the aptitude minimums, research is needed to insure that aptitude standards are current and accurately aligned with job demands. Twenty-five years have passed since the last evaluation.

Skinner, J. (2007). *Occupational learning difficulty for establishing aptitude requirements.* San Antonio, TX: Operational Technologies Corporation. **E-23**

Weeks, J. (1984). *Occupational learning difficulty: A standard for determining the order of aptitude requirement minimums* (AFHRL-SR-84-26). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **E-24**

Pre-enlistment Person-Job Match

Classification is the personnel system process for assigning a person (recruit) to a job (Air Force specialty). The process is complicated and can be accomplished in numerous ways. Some ways are suboptimal in terms of the value (payoff) of classification decisions to the organization (Air Force) and to the person (recruit). Beginning in the 1950s, the Air Force Human Resources Laboratory worked with officials from personnel, recruiting, and training offices to improve classification procedures. Conceptually, the overarching and long-term goal was optimizing person-job matches (PJM) to provide the highest return on investment (recruit productivity) to the Air Force. In practice, the collaborative results were suboptimal for practical reasons, but the research demonstrated the revised methods were much better than those the Air Force had been using.

Research focused on two stages in the classification process: pre-enlistment PJM and post-enlistment PJM. The pre-enlistment PJM occurs when the recruit is either assigned to a specific AFS or to one of the broad specialty areas (Mechanical, General, Administrative, Electronic) prior to entering service. In post-enlistment PJM, recruits assigned to one of the four specialty areas are given final job assignments after entering service and while attending basic training. This summary describes research begun in 1972 on the pre-enlistment process.

The Air Force was using a manual process called PROMIS (Procurement Management Information System), which had been implemented in 1971. Recruiters would call a central location to see what jobs were available to a recruit being processed at the MEPS. The manual process resulted only in filling each AFS or aptitude area with a recruit meeting minimal requirements. The process was slow and so over-taxed with recruiters calling in that the telephone system often shut down entirely. Air Force officials were concerned about their competitiveness with the Army which had a more advanced and automated job reservation system.

To enhance PROMIS, AFHRL developed and demonstrated a computerized job reservation system, which was similar to the Army's, and worked like an airline reservation system. The upgraded system allowed assignments via a computer network in real-time. After numerous cost analyses, terminals were installed at the MEPS, where liaison NCOs (LNCO) would look at available job openings to offer recruits. The "prestige value" was considered since potential recruits would see a real-time assignment to an AFS or aptitude area. However, the system was clearly still suboptimal in terms of payoff to the Air Force. The jobs were filled, but there was no consideration of the characteristics of jobs or people, beyond their basic qualifications, which would improve the value of assigning the recruits to alternate jobs.

The next milestone in Air Force classification was a system called Advanced Personnel Data System's – Procurement Management Information System (APDS-PROMIS). ADPS-PROMIS reflected a significant step forward toward optimized assignments. AFHRL developed a system which considered numerous job properties and person

characteristics for predicting the payoff or worth to the Air Force of assigning the recruit to all open jobs for which he/she was eligible. The algorithm for computing payoff values addressed job fill quotas and the rate at which the jobs needed to be filled, minority representation, job aptitude difficulty, predicted technical school success, and the recruit's job preferences. Developing the algorithm made use of a policy decision procedure developed by AFHRL called policy specifying. Policy specifying allowed components to be differentially weighted in the algorithm to meet policy-makers' judgments about their relative importance in computing a decision index (DI) for alternate jobs to be offered to each recruit. Fifteen AFSs with the highest DI values were offered to each potential recruit.

An important feature of the DI was that it considered not only the actual recruit waiting for a job but also a forecast of the characteristics of future recruits eligible for each job. The ideal situation for optimizing classification is to process all applicants at the same time (called batch processing). However, in the real world of pre-enlistment PJM, each recruit had to be processed sequentially (one recruit at a time) through APDS-PROMIS. Thus, it was not possible to use algorithms that would have maximized payoff values under batch processing conditions (the best solution). However, the DI was a close approximation for sequential assignments and was a major advancement in classification procedures at the time. The APDS-PROMIS was implemented in 1976. Analyses of monthly assignment data showed that 36.5% to 59.9% of recruits selected one of the top three AFSs offered by the Air Force, and of those recruits stating a preference for a particular MAGE area, 37.9% were assigned to that preference. The APDS-PROMIS system was clearly more advanced than the other Services' systems, efficient and cost effective, and made assignments with demonstrated value to the Air Force and the recruit.

The success of the project was largely attributed to the collaborative nature of the effort among AFHRL scientists and officials from operational offices who worked hand-in-hand during the development phase. However, over time, the PJM working group, which established the system, met less frequently and then not at all. Interested personnel migrated to new jobs and finally, there was no one left to oversee maintenance and improvements. The system gradually fell into disuse due to lack of strong proponents in the operational communities. Today, classification actions are handled by LNCOs at the MEPS using procedures similar to the airline reservation system from the early 1970s, in the days before APDS/PROMIS. Recruits are typically shown job openings for which they are physically and mentally qualified. The process results in about 60% of assignments through the Guaranteed Training Enlistment Program (GTEP).

The AFHRL research foundation and the Air Force operational experience with ADPS-PROMIS clearly showed that a pre-enlistment PJM system that increases classification effectiveness and productivity payoffs is feasible and doable. However, it requires a force management environment in which personnel, recruiting, and training officials can commit to and support classification goals beyond job fill.

Piña, M., Jr. (2007). *Pre and post enlistment person job match (PJM)*. San Antonio, TX: Operational Technologies Corporation. **E-26**

Post-enlistment Person-Job Match

Post-enlistment classification decisions are made during Basic Military Training (BMT) at Lackland AFB. When airmen are processed at the MEPS, some are given an assignment in one of the four broad specialty areas (Mechanical, Administration, General, Electronic). At the end of BMT, those airmen are assigned to an Air Force Specialty (AFS) within the specialty area.

The Air Force Human Resources Laboratory (AFHRL) worked with training officials on two occasions to improve the post-enlistment classification system, first in the 1960s and then in the 1980s. Both efforts showed that the value of assignments to the Air Force in terms of the overall productivity of airmen could be improved.

The 1960s effort was to develop a personnel classification process to help managers make assignment decisions that would assign/fit the best person to an available AFS. What was needed was a computed value of worth (payoff) for assigning a particular person to a particular job and a system optimization algorithm (called the transportation algorithm) that assigned people to specialties so that the total payoff was maximized. The AFHRL demonstrated that assignments could be made which were better than those resulting from the process ongoing in the training command. For awhile, AFHRL used their computer capabilities to make operational assignment decisions with a transportation algorithm and then provided the results (called assignment recommendations) to Air Training Command (ATC) personnel at Lackland AFB, who processed the job assignments for trainees. This procedure substantially improved the quality of assignments for the Air Force and for individual trainees but there were two problems. The ATC personnel never reached the point of taking over the assignment algorithm, because of their limited computer and programming capabilities at the time. The AFHRL was forced to stop the operational assignment process because of issues with using funds designated for research on an operations and maintenance (O&M) effort.

In the late 1970s and early 1980s, AFHRL again worked on the post-enlistment classification problem. By then, BMT was using the Processing and Classification of Enlistees (PACE) system, a simplistic, non-optimal sorting routine for post-enlistment person-job matches. The system was run weekly to assign jobs to airmen as they graduated from BMT. It combined a trainee file, a job quota file, and an AFS prerequisite file to generate a record for each job for which each person in a week group was eligible. The fields in the records were used to sort, the first sort field being the most important for assignment consideration, the second record field being the next in importance for assignment, etc. Essentially, the PACE was largely driven by short-term management priorities like unfilled technical training class seats. The potential and background of trainees received little consideration. A major shortcoming of the sorting sequence was that highly qualified trainees may be assigned to low-skill jobs, sometimes leaving only the less-qualified trainees for the more difficult jobs.

The AFHRL proposed to enhance PACE with a system that reflected Air Force classification policy, optimally classified personnel based on that policy, and was responsive and easy to use. A decision-modeling technique was developed to help classification experts from the personnel and training communities define their policy for the post-enlistment system. Two (often competing) components were represented: (1) efficiency (time, money, fill priority) to meet the short-term goals of the training system, and (2) effectiveness (aptitude, vocational interest, trainability) to meet longer-term goals of airmen performance, retention, and readiness. The payoff algorithm for an enhanced PACE system and the data/variables (X) and mathematical functions (F) that were selected to represent the experts' policy are illustrated below.
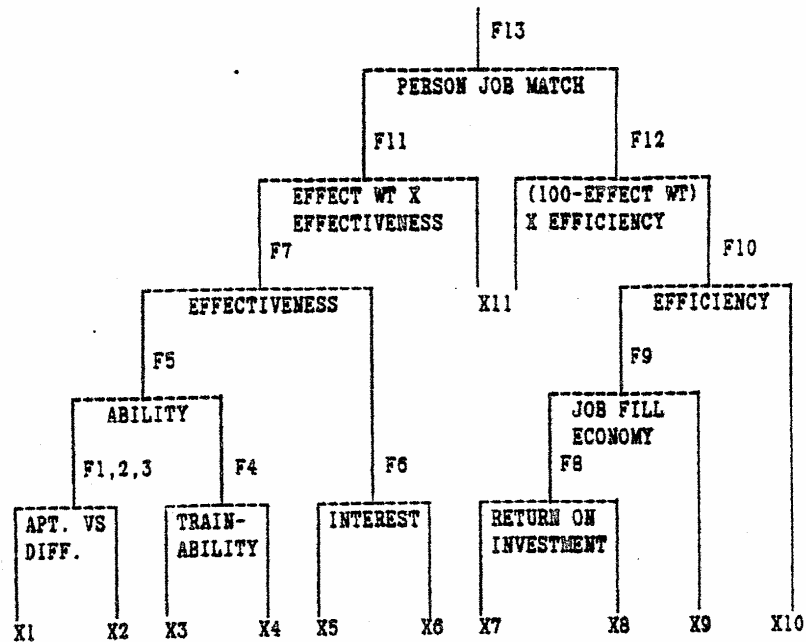
Although the system responded to management's requirements and would have greatly improved the payoff of person-job matches, it was not implemented on a long-term basis to replace the non-optimal sorting process in PACE. Current Air Force managers with an interest in improving post-enlistment classification effectiveness should look to this research foundation as a starting point. Today the Air Force uses an automated procedure called the job spin process to assign a recruit to a specific training class seat. The weighted algorithm incorporates both effectiveness and efficiency measures that are similar to but less comprehensive than the enhanced PACE developed by AFHRL.

Piña, M., Jr. (2007). *Pre and post enlistment person job match (PJM)*. San Antonio, TX: Operational Technologies Corporation. **E-26**

Decision Index:
Simulating Batch Assignments with a Sequential Process

Personnel managers responsible for the distribution and classification of recruits at the Military Entrance Processing Stations (MEPS) and at Basic Military Training (BMT) need to make decisions about the estimated worth of each recruit in various Air Force specialties to maximize the overall effectiveness of the Air Force. The Air Force Human Resources Laboratory (AFHRL) developed and verified the utility of a technique to aid in arriving at such decisions. The technique involved providing a Decision Index for each recruit in each proposed job assignment.

Recruits may be assigned to Air Force specialties (AFSs) using either sequential or batch processes. Sequential methods assign recruits to jobs on a "first come, first served" basis. Batch processes involve assigning a group of recruits to a group of jobs. It is well known and well documented that an organization using a batch process for job assignments will achieve more optimal person-job matches with higher payoff than an organization making assignments one at a time with a sequential process. The "payoff" from any person-job match is the inherent value, utility, or worth associated with placing a recruit in a particular job. The "payoff" can be calculated in a variety of ways including predicted technical school grade, probability of completing a term of enlistment, and/or cost measures. The problem arises that, for organizations like the Air Force, the sequential process is preferred because it is more convenient and tractable for recruiters and assignment counselors to handle one recruit at a time.

The problem addressed by AFHRL in developing the Decision Index was providing the assignment counselors with information which essentially allowed a batch processing outcome to be approximated within an operational sequential assignment framework. Conceptually, the Decision Index can be understood by considering the expression that "the past is prologue to the future." Using historical information about past recruits, it is possible to accurately estimate the quality and number of recruits that a recruiter will see in the future. The Decision Index incorporated the historical information to help recruiters make classification decisions about each individual recruit. As they were processed one at a time, the recruit's payoff value was provided for all jobs, and the Decision Index considered the number and quality (higher or lower) of future applicants. Recruiters could then use the Decision Index to take into account the likelihood (low or high) future recruits of similar or better ability or "payoff" would be available to fill current and future job openings and class seats. The payoff for each individual recruit for all jobs, using the Decision Index method, reflected the value of job placements under batch processing conditions. The Decision Index was used operationally in the person-job match system in the 1970s and 1980s which was called the Air Force Advanced Personnel Data System's - Procurement Management Information System (APDS-PROMIS).

Follow on research was completed to determine the optimality of the Decision Index when used in a sequential manner. Simulations were conducted with varying batch sizes, personnel rejection rates, person-to-job ratios, and personnel payoff distributions.

Comparisons were made with optimal (true batch) and minimal (random) assignment solutions. The finding was that sequential assignments with a Decision Index could attain approximately 92% of the utility of an optimal batch solution for all conditions. These results quantified the value of the Decision Index method for organizations, like the military Services, that must deal on a day-to-day basis with the challenge of assigning large numbers of recruits to many different jobs. The research foundation for the Decision Index as well as prior operational success with the methodology offers an opportunity for the Air Force to improve its current assignment procedures to approach more optimal person-job matches and make better use of incoming recruits capabilities.

Grobman, J.H., Alley, W.E., & Pettit, R.S. (1995). *The optimality of sequential personnel assignments using a Decision Index* (AL/HR-TP-1995-0026). Brooks AFB, TX: Human Resources Directorate, Armstrong Laboratory. **E-38**

Ward, J.H., Jr. (1959). *Use of a Decision Index in assigning Air Force personnel* (WADC-TN-59-38). Lackland AFB, TX: Personnel Laboratory. **E-39**

Differential Assignment Potential in the ASVAB

The use of aptitude tests for military personnel selection has a well-documented history. In comparison, relatively little attention has been devoted to the process of classification which involves allocating applicants to two or more jobs based on differences in the utility of alternative assignments. Studies completed in the military Services have yielded equivocal results from two research streams about the value of the ASVAB for differential assignment decisions. Each of the Services still creates from four to 10 classification composites for qualifying entrants for entry into specific military specialties.

One stream of research evolved from analyses of the structure of cognitive abilities. In widely-replicated findings not only with the ASVAB subtests but also with other multiple aptitude batteries, the scores on the tests were shown to exhibit positive manifold. That is, scores on virtually any well-constructed and reliable measure of cognitive ability will be positively correlated with scores on other cognitive measures. The examinees who tend to do well on vocabulary items will also tend to do well on paragraph comprehension and spatial items. The utility of a multiple aptitude battery for personnel selection without positive manifold would be highly suspect. Classification composites derived from the ASVAB by the military Services are often highly intercorrelated due to the underlying positive manifold. Further, analyses of ASVAB test structure, as well as that of other well-known multiple aptitude batteries, reveal a principal factor measuring general mental ability (also called psychometric g) or working memory. In the case of the ASVAB, specific ability factors have also been identified. The specific ability factors add marginally to the general ability component in predicting technical training and job performance criteria in the Air Force. This research stream has focused principally on the structure of human intelligence. As a consequence, the research interests and methods have not emphasized practical benefits to the Air Force associated with the specific abilities measurement.

A second stream of research, which has been led principally by Army researchers, addressed how different configurations of the ASVAB subtests for classification composite development would produce gains in the utility of the test for differential assignment purposes. The criteria of interest were mean predicted performance in training and on the job. Other researchers documented that adding non-cognitive measures (interests and psychomotor tests) would add to the classification utility of the ASVAB, if the measures were selected to enhance the differential or specific ability content of the test. An Air Force study addressing the controversy of the differential classification value of the ASVAB simulated how recruits could be reassigned to optimize overall job performance based on their ASVAB test scores. The performance gain using optimal assignment over the current assignment baseline was a 1/3 standard deviation unit increase in mean job performance, which was shown to be equivalent to giving recruits an additional 14 months of technical experience. The job performance measure was a work-sample test. Other Air Force studies periodically focused on revising the MAGE composites to improve their predictive effectiveness.

The controversy over the nature of human cognitive and learning abilities and the value of multiple aptitude batteries for differential assignment is an enduring one. It has been the subject of research for more than 100 years as measurement specialists worked to understand and improve mental ability tests for employment decisions. Both research streams have contributed important findings. Recent related projects include the DoD decision to change the content of the ASVAB by adding the Assembling Objects subtest and dropping Numerical Operations and Coding Speed subtests. These content changes have necessitated additional studies of how best to combine the revised ASVAB content into classification composites to assist the Services in assigning recruits to different occupational specialties. A study of the Air Force composites is ongoing in 2007. Other studies are addressing enlisted specialty structures and developing new indices of enlisted job performance. Test content, specialty structures, and performance measures all play a role in assignment decisions.

Alley, W.E., & Teachout, M.S. (1995). *Differential assignment potential in the ASVAB: A simulation of job performance gains* (AL/HR-TP-1995-0006). Brooks AFB, TX: Human Resources Directorate, Armstrong Laboratory. **E-27**

Diaz, T., Ingerick, M., & Lightfoot, M.A. (2004). *Replication of Zeidner, Johnson, and colleagues' method for estimating Army aptitude area (AA) composites* (DTIC Report No. ADA 426 299). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences. **E-28**

Kyllonen, P.C. (2007). *The Learning Abilities Measurement Program (LAMP) 1982-1999*. San Antonio, TX: Operational Technologies Corporation. **E-03**

Ree, M.J., & Carretta, T.R. (2002). g2K. *Human Performance, 15*(1/2), 3-23. **E-29**

Benefits of Selection and Classification

There has been long and continued interest in the topic of estimating benefits of personnel selection in the military and private sector. Much progress has been made to resolve technical issues associated with single-job selection and to extend the procedures toward consideration of the dollar-value utility of using aptitude tests for personnel selection. The research advanced to clearly demonstrate that selection tests contribute large dollar savings to national productivity through improved job performance and reduced attrition.

Research on the benefit estimation in the larger and more inclusive domain of simultaneous selection and classification into multiple-job systems, such as the Air Force uses, has been slower to evolve. In the early 1990s a major review of research concluded with the statement that the topic had been largely unexplored. Most of the research accomplished had been conducted by the military Services. Questions frequently arise in connection with large personnel programs such as those in the DoD concerning which of several alternative interventions might be expected to yield the most benefits. The interventions might include (a) recruiting efforts to expand the military applicant pool in size or quality, (b) employing tests with greater validity to improve the accuracy of selection and assignment decisions, (c) structuring military job specialties to provide more or fewer alternative assignment opportunities, (d) testing a wider diversity of ability domains to improve the differential nature of selection tests, and (e) changing entry standards to accommodate different requirements and recruiting markets.

Procedures for a general solution for estimating selection and classification benefits began to evolve in the 1940s and 1950s with a broad outline of the complexity of the classification problem. Expected criterion performance of personnel selected by means of an aptitude test was characterized as varying according to the number of possible job assignments, the proportion of applicants rejected, and the validity and intercorrelation of the performance estimates (e.g., predicted technical school training scores). Beginning in 1959, researchers made frequent use of what was referred to as the Brogden table for determining classification gains. The tabled entries showed expected performance for 10 levels of applicant rejection rates across 1 to 10 jobs. Although the table was considered a major breakthrough, its value was somewhat limited for the military which typically dealt with classification problems involving more than 10 assignment categories or jobs.

To improve the utility of the table for judging potential benefits of planned enhancements to military selection and classification programs, researchers at AFHRL expanded the table to 500 job categories. The table provides a planning baseline for managers and researchers interested in determining what magnitudes of performance outcomes are feasible to obtain in the DoD environment from changes to aptitude tests, job structures, and recruiting procedures that influence classification gains.

Alley, W.E., & Darby, M.M. (1995). Estimating the benefits of personnel selection and classification: An extension of the Brogden table. *Educational and Psychological Measurement, 55(6)*, 938-958. **E-30**

Joint-Service Classification Research Roadmap

In 1992, the personnel research laboratories for the Air Force, Army, and the Navy sponsored a project to develop a joint-service classification research agenda or roadmap. Many of the research issues identified are current today. The project purpose was to document ways to enhance the overall efficiency of the Services' selection and classification research programs by reducing redundancy and improving inter-service research planning, while ensuring that each Service's priorities in classification research were met. With these goals in mind, the Air Force oversaw a 2-year contract effort with the Human Resources Research Organization (HumRRO) to develop the classification research roadmap.

The project had six tasks: 1. identify classification research objectives; 2. review classifications tests; 3. review job requirements; 4. review performance criteria; 5. review statistical and validation methodologies; and 6. prepare a roadmap for classification research. Task 1 was accomplished by interviewing scientists and decision-makers from each Service to determine research objectives and to document selection and classification practices. Tasks 2 through 5 were comprehensive and systematic reviews of predictor, job analytic, criterion, and methodological needs of each of the Services. Task 6 was a roadmap for classification research which integrated the findings of earlier tasks into a master research plan.

In Task 6, a final report presented a research agenda for military selection and classification. Recommendations for research activities were organized in seven broad areas: 1. building a Joint-Service policy and forecasting data base; 2. developing new job analysis methodologies; 3. capturing criterion policy; 4. conducting criterion measurement research; 5. conducting predictor-related research; 6. modeling classification decisions; and 7. investigating fairness issues.

A decade later many of the research recommendations from this project have not been fully addressed. The project would serve as a valuable foundation to Air Force managers for designing and updating a research program on military personnel classification. The roadmap project was comprehensive and systematic in its approach and integrated military studies not only from the different Services but also theory, models, and findings from relevant literature in the psychology domain. Many of the research needs identified are still current or would require few additions to bring them up-to-date.

Campbell, J.P., Russell, T.L., & Knapp, D.J. (1994). *Roadmap: An agenda for joint-service classification research* (AL/HR-TP-1994-0003). Brooks AFB, TX: Human Resources Directorate, Armstrong Laboratory. **E-31**

Skinner, J. (2007). *Joint-service classification research roadmap*. San Antonio, TX: Operational Technologies Corporation. **E-32**

Enlisted Trends

Characteristics of airmen have been tracked and analyzed since the Air Force was established in 1947. The Air Force needs to have information about their airmen to be able to set and meet recruitment standards and ensure that the people who are selected are capable of successfully supporting the mission. They are interested in the aptitude levels of those selected for service and their demographics such as age, educational level, gender, ethnicity, marital status, and region of the country they come from. The Air Force Human Resources Laboratory (AFHRL) was responsible for tracking the trends of the enlisted force until 1974 when the responsibility for tracking and publishing enlistment trends was moved to the Office of the Under Secretary of Defense (Personnel and Readiness). Several trends in the enlisted force are noteworthy since they have had significant impacts on the composition of the Air Force. These include changes resulting from the All Volunteer Force (AVF), the role of women, and minority recruiting.

The AVF began in 1973 after several decades of the draft. A comprehensive study known as the Gates Commission looked at the potential impact on the Services of an AVF and surmised that it would enhance the efficiency and dignity of the Armed Forces. Those opposed to AVF believed that the military personnel would be less qualified, primarily come from the lower economic classes, and have an overrepresentation of Black personnel. A number of studies were conducted beginning in 1970 at AFHRL to determine what impact the AVF would have on the Air Force and the Military. These early studies predicted new accessions with lower aptitude levels, lower educational levels, and an increase in the proportion of black enlistees. There was also some concern that the Air Force would skim off the best of the manpower pool. Studies conducted in the early years following the AVF showed that the concerns were unfounded. The Air Force did not take the best of the manpower pool, but enlisted those from the "central aptitude spectrum" and returned many of them to the civilian workforce trained to do many jobs. The Air Force had no problems in recruiting qualified enlisted personnel and even raised their selection standards in 1975. Overall, the number of Black enlistees remained about equal to the proportion in the population.

Another significant trend that has changed the makeup of the Air Force is the increasing role that women have played in the mission. In the 1950's, almost all women were placed into clerical and administrative jobs. When the Air Force Qualifying Test became more heavily weighted with mechanical information in 1956, the Armed Forces Women's Selection Test (AFWST) was developed to reduce bias against women. As time passed, more women began to enter the mechanical and electronics fields as their aptitude scores in these areas increased. By the mid-1970's, recruiters had the goal of enlisting 25% of the women in the electronics career fields and 25% in the mechanical career fields. Female enlistments have risen over the years and today the Air Force has the largest proportion of female recruits of any of the Services, partly because almost all job specialties are open to women. In 1970 only 5% of the accessions were female; but by 2004, 22% of the accessions were female.

As mentioned earlier, there was concern that the AVF would result in an overrepresentation of Black personnel in the Services, but the Air Force continued to

recruit Black accessions that were proportionate to or slightly higher than the population representation. A study in 1971 reported that most Black enlistees came from the South and Southwest and that 64% of them fell into the Category IV aptitude level. By 1974, the aptitude scores had risen to an average of 50 from an average of 43 in 1972. Overall, by 2004, Air Force accessions were 22% minority. The minority population, especially the Hispanic population, continues to grow as reported by the U. S. Census Bureau, and the Military may see an upward trend in minority accessions in the future.

The Population Representation Reports that the DoD began publishing in 1974 show how the aptitudes and demographics of military personnel have changed over the years. The data are provided for enlisted personnel, officers and reservists. Quality personnel are defined as those enlistees who have AFQT scores (scores derived from the ASVAB) at the 50th percentile or higher and who have a high school diploma. In 1987, the DoD implemented a Tier System of educational level with Tier 1 as the highest. All those in Tier I have a high school diploma, an adult diploma, or have at least 15 hours of college credit. In 2004, 99% of Air Force accessions were in Tier 1. Aptitude levels were also very high. In 2004, 82% of all Air Force recruits scored at the 50th percentile or higher on their AFQT scores. Overall, 81% of Air Force recruits met the requirements of quality personnel.

Tables taken from the 2004 Population Representation Report showing data for end strength, high quality non-prior service accessions, non-prior service accessions with high school diplomas, gender representation, ethnicity representation, and marital status can be found in the Appendix to this paper.

Since its inception in 1948, the Air Force has been on a journey toward increased enlistment quality. Even though the end strength for the Air Force has been reduced, the quality of the force has increased. In 2004, 81% of enlistees met the criteria for high quality personnel, the number of women had increased and they were performing jobs across the spectrum of specialties, Black representation was on par with the representation in the population, and the Air Force was able to sustain its mission under the auspices of an All Volunteer Force.

Population representation in the military services, Fiscal Year 2004 (2006).
    Washington, D.C.: Directorate of Accession Policy, Office of the Under Secretary
    (Personnel and Readiness). **E-33**
    http://www/dod.mil/prhome/poprep2004/introduction/index.html

Thompson, N.A. (2007). *Tracking trends in the enlisted force.* San Antonio, TX:
    Operational Technologies Corporation. **E-14**

Appendix
Data Taken From the 2004 Population Representation Report Published by the
Office of the Under Secretary (Personnel and Readiness)
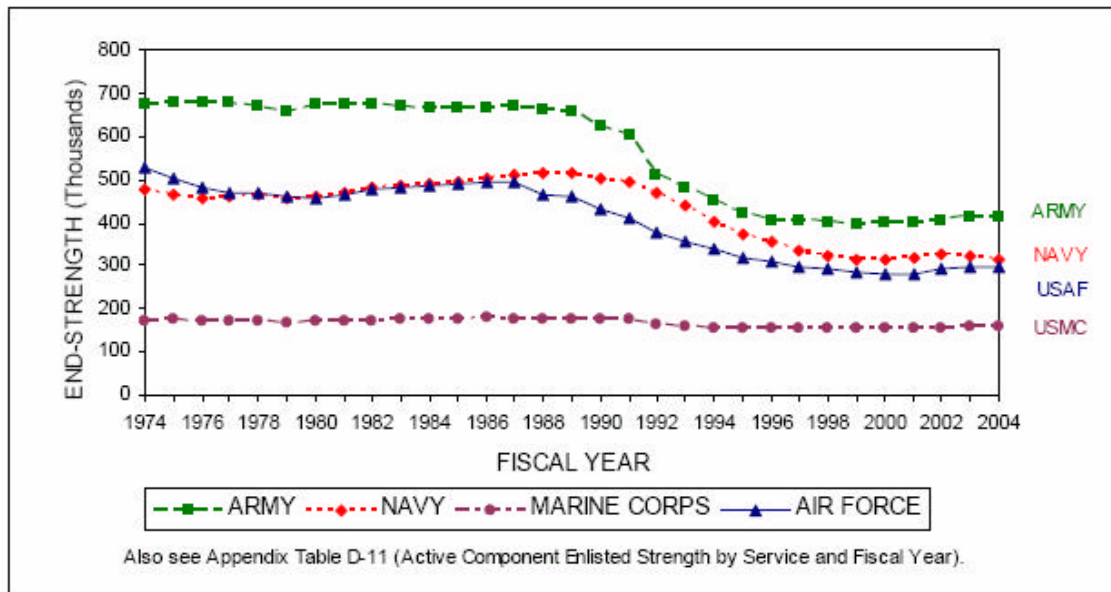

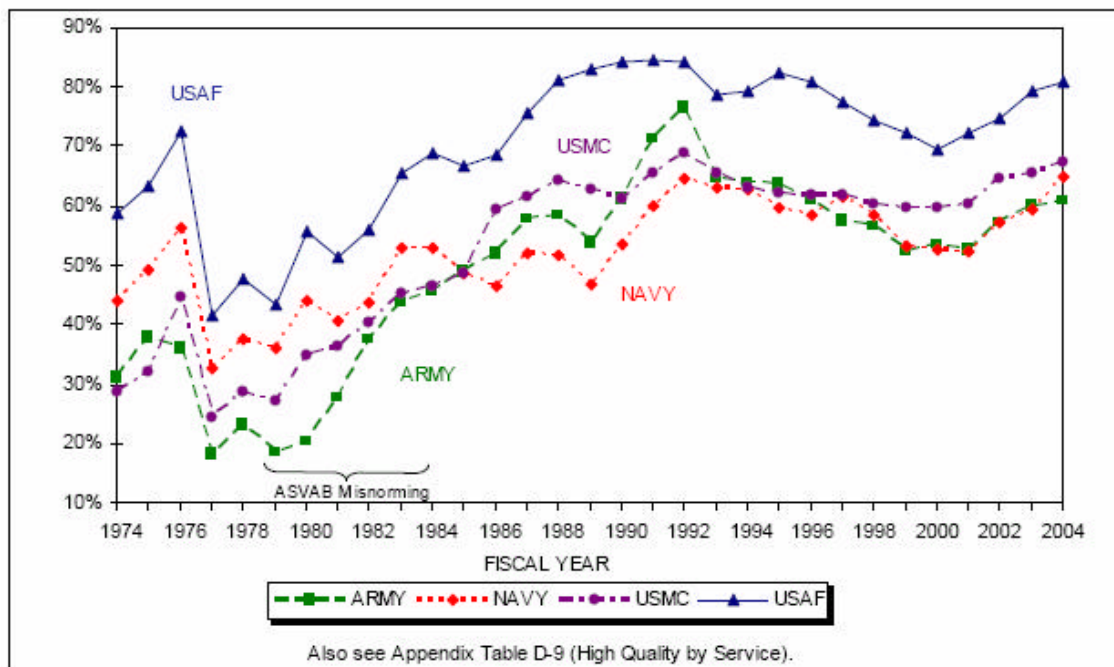**Figure 3.1.** Active Component enlisted force end-strength, by Service, FYs 1974–2004.

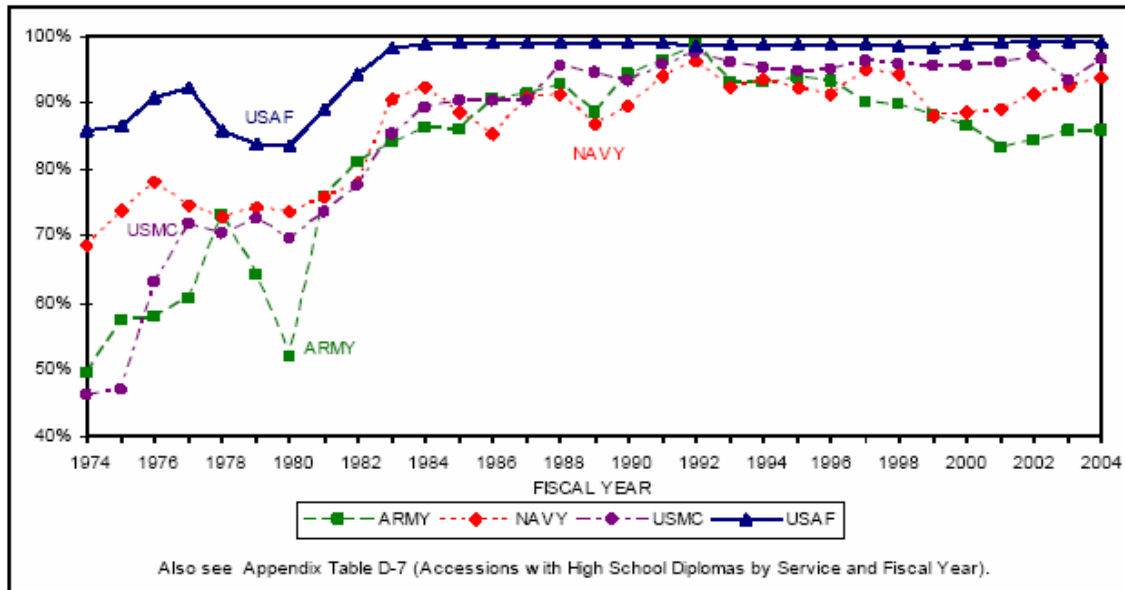**Figure 2.8.** Percentage of high-quality NPS accessions, FYs 1974–2004.

Also see Appendix Table D-7 (Accessions with High School Diplomas by Service and Fiscal Year).

Figure 2.4. Active Component NPS accessions with high school diplomas, FYs 1974–2004.

Also see Appendix Table D-13 (Female Enlisted Members by Service and Fiscal Year).
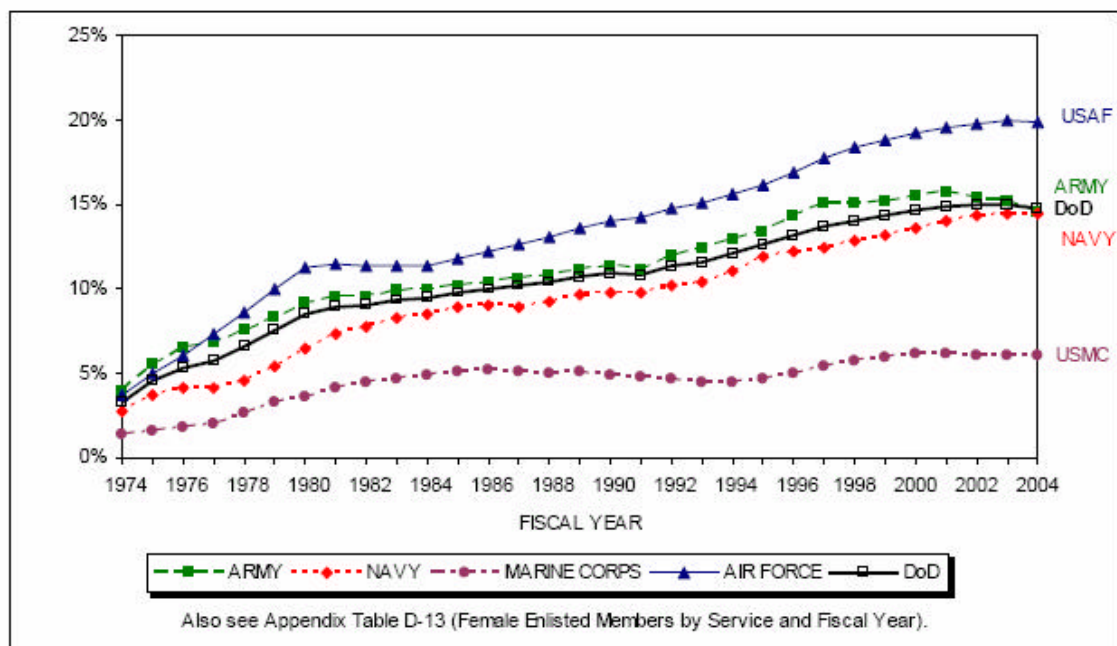
Figure 3.3. Women as a percentage of Active Component enlisted members, by Service, FYs 1974–2004.
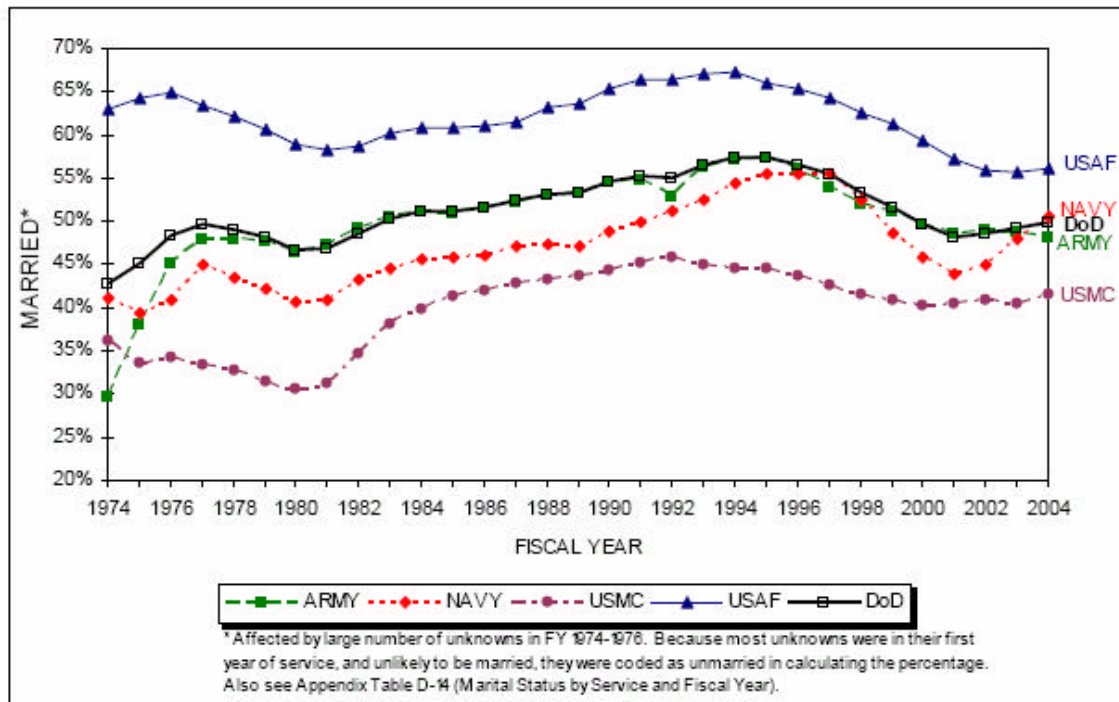
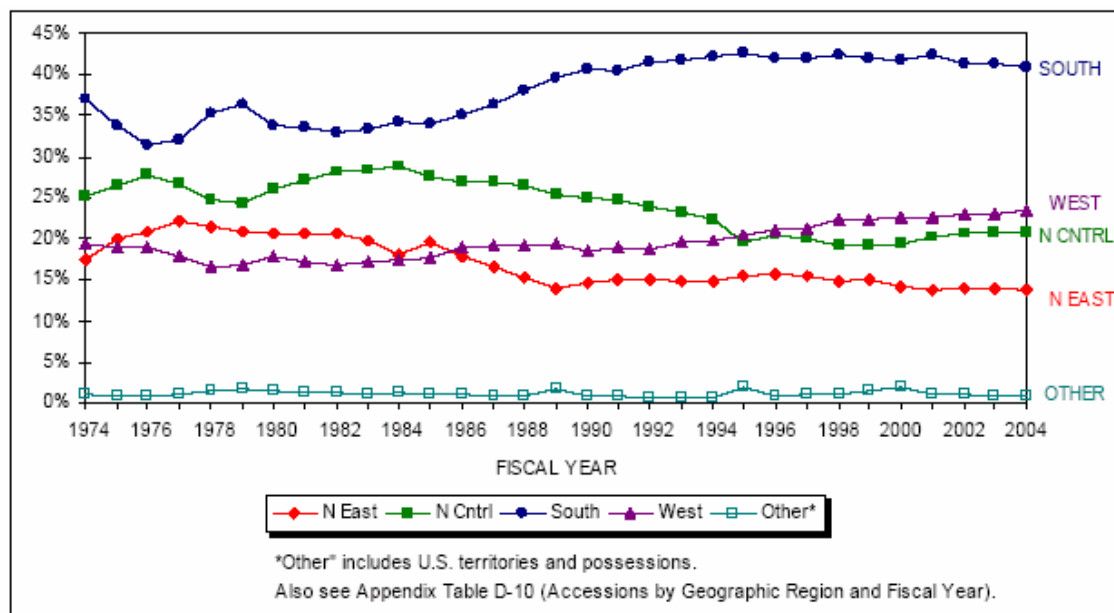**Figure 3.4.** Percentage of Active Component enlisted members who were married, by Service FYs 1974–2004.

Figure 2.9. NPS accessions by geographic region, FYs 1974–2004.

First-Term Attrition

In 1997, the GAO reported that more than 30 percent of recruits across Services leave before the end of their first term. The Services lose a substantial investment in training, time, equipment, and related expenses and must increase accessions to replace the personnel who fail to complete their initial enlistment. The GAO concluded that if the Services were to actually reach their goals in reducing attrition, they would realize immediate short-term annual savings ranging from $5 million to $39 million. First term attrition is a significant personnel problem in all of the military Services.

There is a substantial research foundation on identifying individual characteristics and providing profiles of recruits likely to attrite in their first term of service. Studies show that educational level, aptitude, age, marital status and gender are factors related to successful completion of an obligated term.

By far, the best predictor of attrition is educational level. Attrition rates are consistently higher for non-high school graduates than for high school graduates. The attrition behavior of recruits with alternative educational credentials like the GED more closely resembles that of non-high school graduates than that of high school graduates. High school graduates are judged to be more adaptable to military training and the personal characteristics of maturity, perseverance, and tolerance for rules that contribute to high school completion are also seen to be linked to their likelihood of successfully completing a contracted military service obligation.

Aptitude is related to attrition but is not as strong a predictor as educational level. Recruits in higher score categories of the Armed Forces Qualification Test (AFQT) of the ASVAB are more likely to complete skills training and their first term of service than those in lower aptitude categories.

Age is predictive of attrition, although the relationship is relatively weak. Loss rates tend to be highest for 17-year olds, lowest for 18- and 19-year olds, and moderately high for recruits 21-years old and older. There is some evidence that younger recruits tend to leave for behavioral causes, while older recruits are more likely to leave for medical reasons. In addition, some research on older recruits has suggested that a long gap between high school completion and entry into military services may signify social adjustment problems.

Studies have consistently found that married enlistees are more likely to leave service early than single enlistees. The finding holds for both males and females.

Research on gender differences in attrition indicate that women are more likely to leave the military before the end of their first term than are men. Females tend to leave for medically-related reasons including pregnancy, and males are more likely to leave for disciplinary-related reasons.

Research is less plentiful and/or findings have been equivocal on other personal characteristics like race, geographical region of enlistment, and presence or absence of dependents. Increasing research attention is being given to moral character. Military recruits with offense histories related to criminal behavior and substance abuse are more likely to become premature losses, but the relationship is moderated by accession policies requiring waiver holders to meet higher quality standards on educational level and aptitude factors. The role of matching vocational interests to job requirements, a process known to enhance job satisfaction, has also been explored.

In addition to personal characteristics, organizational and situational influences on attrition behavior have been examined. Studies have found differences in premature attrition rates between Services. Within the Air Force, differences have been detected in attrition rates among occupational specialties. Participation in the Delayed Entry Program is associated with bwer loss rates in the first term. Management and/or administrative policies have also been shown to be significant factors in managing attrition rates. Several studies pointed to the importance of realistic job previews in controlling attrition rates.

The DoD responded to research findings on attrition by placing a strong emphasis on recruiting quality recruits --- those with high aptitudes and high school diplomas. Nevertheless, attrition rates have remained high. Management options for controlling attrition are complicated by the fact that loss rates are not uniform across the time from entry into service until separation. Further, the major contributing factors are not the same for recruits who leave at different points in the first three years of service. The shift in military organizational climate since the beginning of the All-Volunteer Force policy has also been implicated as a factor contributing to attrition despite improvements in amenities offered to recruits. Some researchers have conjectured that the positive implications of a high school diploma for motivation and discipline have changed in recent decades, possibly due to declines in educational standards at the high school level.

Monitoring and tracking attrition is a high priority for all the Services to determine why attrition rates remain steady, even while trend studies show improvements in the quality of recruits during the past several decades. In 2007 a large-scale investigation of attrition in the Air Force was begun under the sponsorship of HQ AFMPC/DPST.

Finstuen, K., & Alley, W.E. (1983). *Occupational correlates of first term enlisted tenure* (AFHRL-TR-82-36). Brooks AFB, TX: Air Force Human Resources Laboratory. **E-34**

Laurence, J.H., Naughton, J., & Harris, D.A. (1996). *Attrition revisited: Identifying the problem and its solutions* (ARI Research Note 96-20). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences. **E-35**
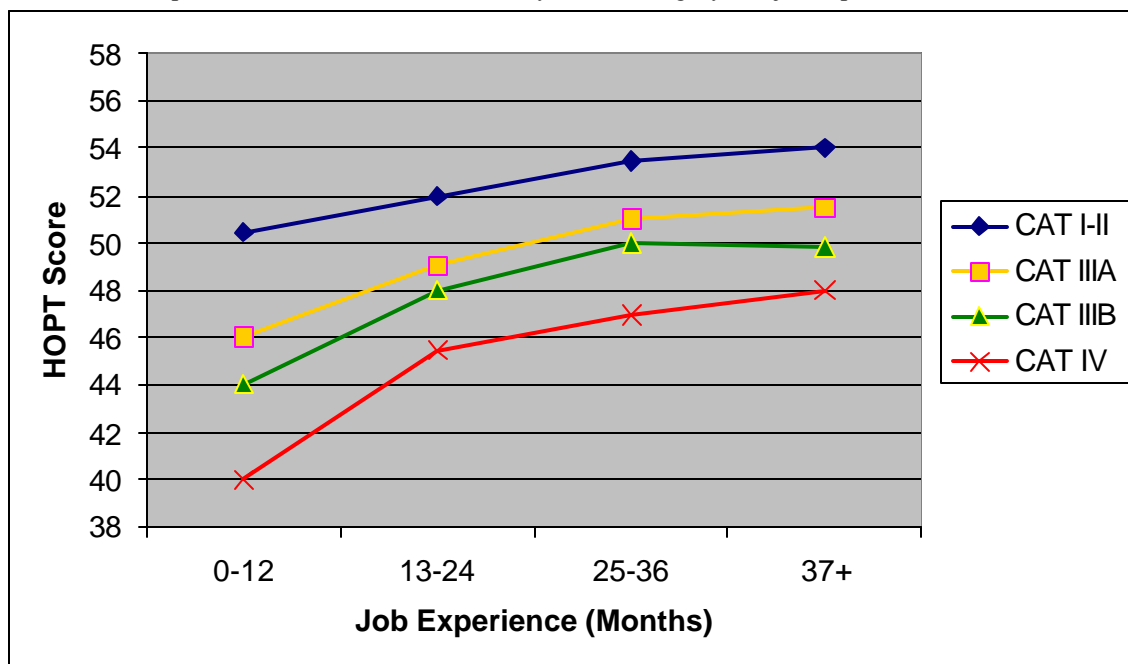
Job Performance Measurement (JPM)

Joint-Service JPM/Enlistment Standards Project

In 1981, the Services launched a pioneering research program to develop measures of job performance so that for the first time, enlistment standards could be linked, at least on a limited basis, to performance on the job. The project was directed by the Office of the Assistant Secretary of Defense (OASD) for Force Management and Personnel. The National Academy of Sciences (NAS) provided technical review. Each of the Services developed programs of performance measurement research. Policy makers in Congress and the DoD mandated the effort requiring the Services to establish an empirical relationship between the ASVAB and actual job performance. Historically, the relationship between the ASVAB and technical training grades has been used as a basis for selection and classification decisions.

Hands-on work sample tests were identified as the primary measure of job performance and were a common feature of the Services' research programs. Hands-on tests required job incumbents to actually perform a task in the workplace with the tools and equipment used on the job. Elements of correct performance were scored by trained observers and task scores were obtained. The validity of the Armed Forces Qualification Test (AFQT) for predicting hands-on performance measures were reported to the House Committee on Appropriations in 1989. Validities were reported for 23 occupational specialties, eight of which were Air Force specialties. While the correlations were generally lower than those obtained using training grades as criteria, the overall results indicate that the AFQT has a positive relationship with hands-on performance. Other analyses showed that hands-on scores increase with level of experience within AFQT score ranges (see Figure below).

Mean hands-on performance test (HOPT) scores by AFQT category and job experience.

The NAS concluded that the JPM Project succeeded in demonstrating that hands-on measures of job performance could be developed for a wide range of military jobs and that the ASVAB predicts these measures with a useful degree of validity. They pointed out that a remaining task was to use the results to link enlistment standards to job performance. Work continued to develop methods for linking recruit quality requirements and job performance data. As a result of this work, OASD instituted benchmarks stating that sixty percent or greater of all recruits must be in Categories I – IIIA (i.e., at or above the 50th percentile), and 90% must be high school graduates. In addition, an overall DoD model as well as individual Service models were developed to capitalize on cost/performance trade-offs. However, each of the Services retained occupational classification standards based on aptitude and training performance relationships.

Several reasons have been given against using measures of on-the-job performance in setting enlistment standards. Compared to training performance, more factors influence an individual's job performance other than their cognitive ability. These factors include organizational, team, unit and individual variables. Differences in operational requirements, leadership, and situational variables; differences in opportunities to perform tasks; and differences in individual recruit's motivation, satisfaction and commitment have been identified as potentially affecting level of job performance. These differences affect individual performance to a greater extent in the workplace, while training grades are more a function of the ability of recruits. Further, the non-ability factors tend to obscure the relationship between a cognitive ability predictor and a performance criterion. Hence, the relationship between a predictor (i.e., ASVAB) and measures of job performance (i.e., HOPT) is usually lower, as there is an increase in time between measuring the predictor (i.e., ASVAB score) and the performance criterion. Numerous studies show that the ASVAB is an excellent predictor for what is needed for an airman's first job. Since there is a substantial relationship between training grades and job performance, training grades may be sufficient for the interim for setting selection and classification standards.

Existing operational measures such as airman performance reports and promotion test results are inadequate for measuring individual performance and lack required reliability and validity for selection and classification purposes. Developing new measures of job performance as was accomplished in the JPM project, however, is cost prohibitive. In each of the eight AFSs examined in the Air Force project, research costs exceeded $1 million. Alternate approaches for job performance measurement continue to be of interest. As of the time of this writing in 2007, HQ AFPC/DPST is sponsoring a contract effort to explore the feasibility of developing inexpensive individual measures of job performance from archived personnel data maintained by the Air Force Personnel Center.

Teachout, M.S. (2007). *The Joint-Service Job Performance Measurement/Enlistment Standards Project and the Air Force Job Performance Measurement Project: A summary of key results.* San Antonio, TX: Operational Technologies Corporation. **E-21**

Air Force Job Performance Measurement (JPM) Project

The Air Force Human Resources Laboratory (AFHRL) completed a large-scale effort to develop a measurement approach for systematically obtaining job performance data. Program managers in the manpower, personnel and training (MPT) communities requested the project for evaluation selection and training programs. The project was underway when an additional requirement arose, the Congressional mandate in 1980 to link military enlistment and classification standards to job performance.

The project goal was development of measurement techniques for collecting reliable, valid and accurate hands-on performance information. Hands-on measures were used as benchmarks against which more affordable, easier to administer measures were evaluated. Work sample tests are the highest fidelity measures of job performance capability and include hands-on performance that require incumbents to display the same behaviors as they would on the job (i.e., perform the tasks using operational equipment, materials and procedures). The resulting Air Force Job Performance Measurement System (JPMS) consisted of Walk-Through Performance Testing, a set of four rating forms, and job knowledge tests.

The Walk-through Performance Test (WTPT) combined the observation of hands-on performance testing (HOPT) with interview testing. The hands-on component of the WTPT was a traditional hands-on work sample test designed to measure proficiency on selected job tasks. Participants were asked to actually perform specific tasks in order to demonstrate their proficiency. The interview component used a show-and-tell approach, where participants described, rather than performed, the step-by-step procedures they would do to successfully perform a task. Since many tasks would be too time-consuming, costly, or dangerous to measure using the hands-on method, the interview method was developed. Results for eight Air Force specialties showed positive relationships between the scores obtained by airmen on the Armed Forces Qualification Test (AFQT) and how well they performed the HOPT (see Table, next page). Further, positive correlations between HOPT scores and interview scores showed that interview testing was a useful alternative to the more costly and time-consuming hands-on testing method. In particular, the interview approach appears promising for the Personnel Specialist, Information Systems Radio Operator, and Air Traffic Control specialties, likely due to the verbal nature of those jobs.

Further analyses were conducted to assess the substitutability of other JPM measures for hands-on tests. As expected, hands-on measures were most strongly related to interview tests and job knowledge tests with performance ratings showing the lowest relationship to hands-on measures. Overall, none of the surrogates were considered interchangeable or substitutable for the hands-on measures.

Correlations between AFQT and HOPT and between HOPT and Interview

ASVAB Composite	Air Force Specialty	AFQT and HOPT*	HOPT and Interview
Mech	Jet Engine Mechanic (AFS 426X2)	.29	.57
Admin	Information Systems Radio Operator (AFS 492X1)	.35	.80
Gen	Air Traffic Control Operator (AFS 272X0)	.16	.81
Elec	Avionic Communications Specialist (AFS 328X0)	.67	.66
Mech	Aerospace Ground Equipment Specialist (AFS 423X5)	.36	.70
Admin	Personnel Specialist (AFS 732X0)	.53	.84
Gen	Aircrew Life Support Specialist (AFS 122X0)	.21	.59
Elec	Precision Measuring Equipment Specialist (AFS 324X0)	.66	.46

* Correlations corrected for aptitude restriction in range.

Several follow-on research studies applied the methodologies to training assessment and evaluation issues. This was a logical transition from the JPM project, since reliable and valid measures of performance are needed to evaluate training programs.

The JPM research made a substantial contribution to understanding the dimensions underlying performance on the job. The major benefit to the Air Force was demonstrating empirically that the airman selection and classification testing program was related to performance beyond technical training in accomplishing actual tasks in the workplace. The military leads the field of applied psychology in job performance measurement. Findings from the Air Force research program were shared with the private sector in professional journals and books.

Teachout, M.S. (2007). *The Joint-Service Job Performance Measurement/Enlistment Standards Project and the Air Force Job Performance Measurement Project: A summary of key results.* San Antonio, TX: Operational Technologies Corporation. **E-21**

Learning Abilities Measurement Program (LAMP)

The Learning Abilities Measurement Program (LAMP) 1982-1999

The Learning Abilities Measurement Program (LAMP) was an Air Force Human Resources Laboratory (AFHRL) project active from 1982 to 1999, designed to improve the military Services' personnel selection and classification (S&C) systems by taking advantage of then new developments in cognitive psychology, computer technology, and psychometrics. The program was initiated with a proposal to the Air Force Office of Scientific Research (AFOSR) in 1982 and was supported with both basic (6.1) and applied (6.2) research funding. The program was inspired by a basic research program on individual differences in cognition supported by the Office of Naval Research (ONR). It also was contemporaneous with somewhat related programs in the Army (Project A, and its successor, Building the Career Force), and the Navy (Enhanced Computer Adaptive Testing [ECAT]), but differed from those programs in its more basic research focus. As such, although initially conceived of as exploiting developments in cognitive psychology to improve enlisted S&C, LAMP touched on many other domains including personality, human factors, cognitive engineering, artificial intelligence, and chronobiology. The program also led to many other applications besides enlisted selection including pilot (and fighter pilot) selection; situational awareness assessment; intelligent tutoring systems; chronotype assessment (morningness-eveningness); automatic item generation; evaluating the effects of drugs, fatigue, and stressors with performance assessment batteries; and others.

Major accomplishments of the program were both applied and theoretical. The primary applied contributions were the development of several aptitude (predictor) batteries, and several criterion training and performance environments. The predictor batteries included information-processing ability batteries (Cognitive Abilities Measurement [CAM] and Advanced Personnel Testing [APT]), a perceptual-motor battery (P-CAM), and a Big 5 personality assessment system (the Trait-Self-Description Inventory [TSD]). The criterion performance environments included several Intelligent Tutoring Systems (ITS) teaching Flight Engineering (FET), Basic Electricity (OHM), and Computer Programming (BRIDGE); a logic gates circuit tutor; a Cessna 172 single engine aircraft simulator, the Basic Flight Instruction Tutoring System (BFITS); the Situational Awareness Flight Training Evaluator [SAFTE]), and a Predator-like Uninhabited Aerial Vehicle (UAV) Simulation. An additional applied accomplishment, remarkable at the time (1982), was the development of an experimental testing facility populated with over 200 testing stations, accommodating 30,000 Air Force basic trainees per year.

Among the major theoretical accomplishments of the program were the development of the cognitive abilities measurement (CAM) framework and measurement system; a learning-skills taxonomy; the alignment of general cognitive ability with working memory capacity; the scaling and development of information-processing speed; the identification of a general temporal processing ability factor underlying performance on both "dynamic spatial" and "psychomotor" tasks; and numerous empirical findings involving associative learning, implicit learning, priming, mathematical models of reaction time tasks, interference effects, and others.

The LAMP was a major AFHRL project, and it continues to influence DoD programs. One of the last major LAMP events before AFHRL closed was the publication of a book on automatic item generation. An advantage of the information-processing approach to measuring abilities, such as identifying working-memory capacity as the "g" factor, is that task specifications are detailed and explicit rather than vague. Automatic item generation has not yet been applied to ASVAB subtests, but a recommendation from a recent ASVAB technical review was to begin research to do so.

Kyllonen, P.C. (2007). *The Learning Abilities Measurement Program (LAMP) 1982-1999.* San Antonio, TX: Operational Technologies Corporation. **E-03**

Advanced Personnel Testing (APT) Battery

Experimental tests developed in the Learning Abilities Measurement Program (LAMP) were transitioned in an Advanced Personnel Testing (APT) battery to address their utility for expanding the ability coverage of the Armed Services Vocational Aptitude Battery (ASVAB). The LAMP tests represented a new direction in individual aptitude testing derived from cognitive psychology. Measurement focused on an examinee's information processing capacity instead of the more traditional assessment of an examinee's knowledge base.

Twelve (12) computer-administered tests were identified from a larger set of LAMP tests for the APT project. The tests measured (1) working memory capacity, the ability to simultaneously store old information and to process new information; (2) declarative/fact learning, the ability to learn new facts; (3) procedural/skill learning, the ability to learn simple, novel rules for classifying facts; and (4) induction, the breadth of procedural knowledge. In information processing theory, these processes are believed to be potential sources of individual differences in cognitive ability. The tests covered three content domains (verbal, quantitative, and spatial) hypothesized to reflect individual differences in relative knowledge (verbal vs. quantitative) and to be independent of differences in declarative or procedural knowledge.

Analyses were conducted with large samples of basic airmen who took the APT battery and had ASVAB scores on record. These trainees were tracked, and their technical training course grades served as criterion variables in a study to compare the predictive validity of APT and ASVAB. The specialties were chosen to represent different aptitude areas (Mechanics, Administrative, General, Electronics) and for their high volumes. The project terminated before all the specialty areas were analyzed, but three technical training courses were analyzed, security police, basic electronics, and aircraft mechanics.

The basic findings were first, there was some evidence for incremental validity of APT over ASVAB, but it was fairly small (delta r = .00, .03, .08, for mechanics, police, and electronics, respectively). There were several analyses, some correcting ASVAB validities for attenuation due to range restriction, but it is not clear that APT validities were corrected for indirect range restriction (so the true increments may be larger). Also, the largest incremental validity, for electronics, was uncorrected (so it may be smaller). A second finding is that APT did show less adverse impact related to race and gender. A third finding was that APT tests were more factorially diverse (less unidimensional) than the ASVAB tests suggesting that APT might provide improvements over ASVAB in classification utility. A fourth finding was that the Fact Learning factor (and to a lesser extent, the Skill Learning factor) reflected an ability not measured in the ASVAB, and one responsible for the incremental validity of APT over ASVAB.

These findings were presented in 1996 to DoD Accession Policy in OASD. The findings presented did not match expectations on the part of some policy-makers that APT could replace ASVAB, or even practically supplement it.

In retrospect, there are three major reasons why this may have been an unfortunate and premature conclusion. First, the finding of lower subtest correlations in APT compared to

ASVAB suggested APT's major contribution could have been increased differential validity (classification potential). Although there was some preliminary work examining classification models with APT, that work was never completed.

A second issue, which may be underappreciated, concerns the importance of method variance. APT was an information-processing battery delivered on computers. During the time of APT data collection, in 1994 and 1995, ASVAB was mostly a paper-and-pencil multiple-choice test (CAT ASVAB went fully operational only in 1996). Technical school grades were also based on paper-and-pencil multiple-choice tests. The impact of method variance was never fully explored.

A third reason for APT's poorer than expected performance could have been that very little time was spent optimizing APT for validity purposes. The test that was evaluated was essentially the same as the test that was initially constructed. No attempt was made to design and evaluate items with better psychometric properties (discrimination, validity), or to select tests to maximize validity, as had been done with the Enhanced Computerized Administered Tests (ECAT) project. It is difficult to estimate the overall effects of these three factors, but it could well be that they are important enough to warrant a reconsideration of APT, or an updated APT-like battery as a supplement to or replacement for the current ASVAB.

Kyllonen, P.C. (2007). *The Learning Abilities Measurement Program (LAMP) 1982-1999*. San Antonio, TX: Operational Technologies Corporation. **E-03**

Sawin, L., Earles, J., Goff, G.N., & Chaiken, S.R. (2001). *Advanced personnel testing project: Final report* (AFRL-HE-AZ-TP-2001-0004). Brooks AFB, TX: Warfighter Training Research Division, Air Force Research Laboratory. **E-36**

Vocational Interests

Vocational Interest Career Examination (VOICE)

Most research on vocational interests has utilized commercially available inventories, most notably the Strong Vocational Interest Blank and the Strong-Campbell Interest Inventory. Because these inventories focus on college-oriented professional occupations, they are not appropriate for the population of Air Force enlisted accessions or for military specialties involving clerical and blue collar work that do not require general education beyond the high school level. The Vocational Interest Career Examination (VOICE) was developed at the Air Force Human Resources Laboratory to improve the quality of vocational guidance and job placement for airmen.

The instrument is a 300-item inventory with a 25-minute administration time. The individual items consist of occupational titles, work tasks, leisure time activities, and desired learning experiences. Airmen indicate relative preferences on a like-indifferent-dislike format. Item responses are converted to two types of scales: basic interest scales and occupational scales. The basic scales measure general interest in various occupational and technical areas. The occupational scales were designed for use in evaluating alternative areas of assignment in specific Air Force occupational clusters.

A concurrent validation study showed that VOICE scales of measured interests effectively distinguished between airmen who reported being satisfied and those who reported being dissatisfied with their career field when tested during their first term of service. Another study showed that interests measured at time of entry into the Air Force accurately predicted an airman's level of satisfaction with their job assignment a year later. Other studies showed that airmen with higher job interests had lower failure rates from technical training, received higher ratings of job performance from their supervisors, and had lower premature attrition rates at 12, 24, and 36 months of service.

Despite the strong research findings, the VOICE was implemented only briefly for operational use at Lackland AFB during the mid-1980s. It was discontinued because Recruiting Service elected to deemphasize job-fit in favor of simpler job-fill procedures in the Processing and Classification of Enlistees (PACE) algorithm. The VOICE was subsequently used in Army research programs to develop AVOICE (Army VOICE) and was one of the foundational instruments reviewed by the Navy to design a web-based vocational interest measurement system called Job Opportunities in the Navy (JOIN). Long term plans in the Navy are to implement the JOIN to help Navy applicants obtain a job match during entry-level assignment which fits their vocational preferences and interests. Air Force managers interested in improving classification efficiency should consider vocational interest measurement as an untapped opportunity.

Alley, W.E., & Matthews, M.D. (1982). The Vocational Interest Career Examination: A description of the instrument and possible applications. *The Journal of Psychology, 112*, 169-193. **E-37**

Promotion Systems

The Weighted Airman Promotion System (WAPS)

The procedure used for enlisted Air Force promotions is of utmost importance to both Air Force management and the airmen who want to be promoted. At its best, the promotion system assesses promotability based on each airman's capabilities and achievements without rater bias. At its worst, the promotion system is a haphazard process without standardization that falls prey to management whims. In 1947, the Air Force first used the decentralized promotion system that had been used by the Army. In 1950, the first regulation was published for promotions and demotions (AFR 39-30), but the location of promotion authority for different enlisted grades changed many times over the years. Selection boards were first mentioned in a 1959 regulation, but major commands and bases were free to develop their own local procedures on how promotion selection boards were run. The decentralized board system and lack of standardization in selection procedures led to numerous complaints from airmen about unequal promotion consideration.

Prior to 1970, the Air Staff and Congress had received many complaints from airmen about the promotion system. The promotion board system was seen to be lacking in three significant areas: (1) airmen eligibles did not know how they ranked in relation to their peers, (2) the nonselected airmen were not advised as to why they were not promoted, and (3) no information was provided to nonselectees regarding what they could do to enhance their future promotion potential. In 1967, a Congressional Special Subcommittee on Enlisted Policy met to address the many complaints on promotion policy across the military. The Air Force was asked to develop a new selection system. In response to the direction of Congress, a highly reliable system called the Weighted Airman Promotion System (WAPS) was introduced for airman eligible for promotion to staff sergeant (E5), technical sergeant (E6), and master sergeant (E7). Airman promotions from E3 to E4 were originally included in WAPS; but that was discontinued in 1971, because the promotion rate for E4 exceeded 90% and it was more cost effective to process the promotions on a fully qualified basis than under WAPS.

WAPS was developed by the Air Force Human Resources Laboratory (AFHRL) as a policy capturing procedure that replicated the decisions that would have been made by a promotion board. The objective was a mathematical model that expressed or "captured" the consensus judgment or "policy" of highly qualified and experienced military personnel about the relative merits of airmen eligible for promotion. Since promotions had been based on the recommendations of promotion boards, the policy-capturing technique identified the optimum variables to be considered in the promotion formula based on the policies that the board members used in ranking airmen for promotion.

AFHRL developed the WAPS by using a three-step approach. First, a promotion board panel of 15 colonels and 16 E8's and E9's was convened. This board rank-ordered a random sample of 2100 E5 airmen for promotion to E6. Each airman's record displayed numerical values for his/her performance on the promotion variables. Each board member was required to review records for all airmen and to make an independent

judgment as to their rank order from most promotable to least promotable. Second, using multiple linear regression techniques, a separate equation was computed for each member of the board. The airmen promotion factors served as predictor variables and the ranks assigned as the criterion. The separate equation for each board member represented the individual's promotion policy. At this stage in the process, there were as many regression equations as there were board members. Third, the multiple equations were reduced to a single consensus equation. To accomplish the reduction, a criterion-grouping technique, referred to as hierarchical grouping, was used to combine the most similar regression equations or promotion policies in an iterative process until there was only one common policy representative of all or the majority of the raters or board members. The final equation provided information, through the size of regression weights for each factor, about the board's judgments concerning the relative importance of each factor to promotion.

Board members ranked the airman sample on several variables. Seven variables were chosen, six of which are still in use today: scores on the Specialty Knowledge Test and Promotion Fitness Examination; scores for seniority based on time in service and time in grade; a score for decorations and medals earned; and, a score for performance ratings. The seventh factor considered in the early development of the WAPS was a promotion board score, but it was found that the inclusion of the board score in the weighted factors system did not influence the ranking of airmen for promotion.

The WAPS has performed well over the last 38 years. After the adoption of WAPS, airman complaints to HQ USAF and congressional inquiries decreased in number. By 1971, the amount of correspondence concerning airman promotion had dropped by 70%. Airman surveys and other feedback revealed airman support and acceptance of the system. The WAPS has been reevaluated over the years with minimum changes to the variables and their weights. The promotion procedures are widely accepted and favorably viewed not only by the enlisted personnel for whom they were designed but also by the personnel managers who are responsible for overseeing and executing enlisted promotion policy and programs.

Shore, C.W. & Gould, R.B. (2004). *Revalidation of WAPS and SNCOPP.* San Antonio, TX: Operational Technologies Corporation. **E-25**

The Senior Non-Commissioned Officer Promotion Program (SNCOPP)

In 1958, the grades of senior non-commissioned officers (NCOs), also called super grades, were established by Public Law 85-422. The grades were Senior Master Sergeant or E-8 and Chief Master Sergeant or E-9, and individuals filling the positions were called senior NCOs. The new grades were a result of House and Senate Armed Services Committee hearings with the intent to reduce high personnel turnover and attract well-qualified personnel in career positions. In addition to providing better career potential, the personnel in the supergrades were meant to perform tasks of higher responsibility with supervisory and management skills. Responsibilities shifted from hands-on skilled technical duties to supervision and management. This policy caused some criticism from the enlisted ranks because an individual could not rise to the higher grades while continuing to perform as a skilled technician. The number of active senior NCOs was capped by Congress at 2 percent of the enlisted force for E-8s and 1 percent of the enlisted force for E-9s.

When the supergrades were established, the Air Force promotion system was evolving from one adopted from the decentralized system that the Army used in 1947, the year the Air Force was designated a separate military Service branch. Changes in promotion policy and procedures during the next 30 years were gradual and reflected moves toward centralization and standardization. In 1959, promotion selection boards were first mentioned by name in a regulation, and a year later more guidance was published defining the composition of the promotion boards.

Then in 1966, a decision was made to centralize E-8 and E-9 promotion selection boards. Central boards had been used on a limited basis for some vacancies prior to that time, but the 1966 decision was significant in that it applied to all promotion selections for grades E-8 and E-9 in all Air Force specialties (AFSs). The need to centralize was forced by gradual decreases in promotion quotas from 1959 to 1964. It became increasingly apparent that the task of promotion selection would fall upon central USAF boards since quotas allocated to most AFSs were too few in number to distribute them equitably to lower organizational levels. Centralization had the benefit of allowing eligible airmen across commands to compete for promotion on equal terms for all vacancies within an AFS.

In 1970, the Air Force studied the possibility of extending the Weighted Airman Promotion System (WAPS) used for lower grade airmen to grades E-8 and E-9. WAPS was developed by the Air Force Human Resources Laboratory (AFHRL) as a policy capturing procedure that replicated the decisions that would have been made by a promotion board. The objective was a mathematical model that expressed or "captured" the consensus judgment or "policy" of highly qualified and experienced military personnel about the relative merits of airmen eligible for promotion. AFHRL was asked to address two questions: 1) can the WAPS be applied to E-8 and E-9 promotions with selection factors weighted as in the system for the lower grades, and 2) if not, can the

same selection factors, optimally weighted, be incorporated in a system which will be suitable for selection of E-8 and E-9 personnel?

The first study in 1970 found that the WAPS variables that had been used in the lower grades did not adequately capture the promotion policy for E-8 and E-9 promotions. In a second research effort, approximately 100 variables in the promotion selection folder such as education, experience, aptitude test scores, performance ratings, and decorations were examined; but no consistent promotion policy was discerned in the policy capturing analyses of panel members' use of the variables in promotion decisions.

A few years later, a third study recommended a dual selection system that combined the most desirable features of the objective weighted factors approach and of the "whole person" scoring approach. The system combined a WAPS-like score based on factors relevant to the selection of E-8s and E-9s with a promotion board score. This concept evolved from discussions with senior NCOs who felt that the current board selection process, while it failed to provide visibility, was a good system for assessing the management potential of candidates for higher grades. Lacking an alternate method of measuring management potential, the concept of a dual promotion system was approved. The first selections using the dual system occurred in 1977.

The dual system was comprised of a weighted factors component similar to WAPS and a promotion board score component obtained from an operational promotion board. A policy capturing approach like the one used for WAPS was applied to the senior airmen to produce a weighted scoring system that could be used for promotion. The factors that were chosen for the system were: United States Air Force Supervisory Examination (USAFSE), Enlisted Performance Report (EPR), Professional Military Education (PME), Decorations (DEC), Time-in-Grade (TIG), Time-in-Service (TIS), and an operational promotion board score.

The original dual promotion system instituted in 1977 has remained largely intact for the past 30 years and continues to meet the needs of promotion to the supergrades. A few changes have been made, all of them minor in nature, and none have altered the essential principles of the dual process. The system combines the best features of two distinctly different approaches to promotion selection decisions: a weighted factors method and subjective board evaluation. The system melds the desired characteristics of objectivity and visibility through the weighted factors component with the judgments of expert panel members through the board score of difficult-to-quantify, but nonetheless important, characteristics for promotability to the supergrades. Properties of the SNCOPP were revalidated in 2004.

Shore, C.W. & Gould, R.B. (2004). *Revalidation of WAPS and SNCOPP*. San Antonio, TX: Operational Technologies Corporation. **E-25**

Job Satisfaction Research Project

The establishment of an All Volunteer Force prompted increased attention to the needs, desires, and attitudes of military personnel. To meet this situation, the Air Force Human Resources Laboratory (AFHRL) initiated a comprehensive job satisfaction research program with the goals of improving the utilization of personnel, retention of qualified airmen, and maintenance of critical skills. Of interest was the impact of occupational factors on job attitudes, productivity, and career decisions. The steps outlined in the research project were (a) to determine the important facets of job satisfaction for Air Force personnel, (b) to examine the relationship between job satisfaction and career decisions, (c) to identify the characteristics of jobs and assignments which produce satisfaction and dissatisfaction, and finally (d) to recommend job and policy changes which would positively effect job satisfaction.

A review of research literature and civilian job satisfaction inventories revealed there were no acceptable instruments available for use in a military environment. To meet the project needs, the Occupational Attitude Inventory (OAI) was developed to measure satisfaction levels among enlisted personnel, primarily those of first-term airmen. The OAI consisted of two major sections: (1) Life History Information, and (2) Occupational Attitude Information. The inventory had 348 items distributed across 35 facets determined to be critical elements or dimensions of military job satisfaction. The facets included Air Force and unit policies and practices, assignment locality, authority, co-worker characteristics, perceived importance of work, pay and benefits, physical work environment, recognition, safety, sufficiency of training, supervision, and value of experience. The need for the large number of facets was supported by a prior study of 97 airman career ladders showing considerable differences in the dimensions of job attitudes operating between and within ladders.

Numerous studies of job attitudes were completed, and findings were obtained on a variety of issues ranging from the role of leisure activities to assignment locality on airman satisfaction. It was found that few sports and leisure activity differences existed between satisfied and dissatisfied airmen when job tenure was taken into account (held constant). However, the importance of one of the OAI facets, characteristics of assignment locality, was demonstrated in several surveys. The facet was the most frequently selected cause of both satisfaction and dissatisfaction. Most preferred locations had large base and civilian community populations, were closer to the ocean and desert, and had a 2-year college readily available. Distributions of even the least preferred locations indicated that significant numbers of airmen in varied specialties did like the locations. Findings had implications for initial assignment decisions and permanent change of station (PCS) policies in staffing least preferred bases and controlling PCS turbulence.

Administration of the OAI and occupational surveys showed that most airmen found their jobs interesting and reported their talents and training were well utilized. However, there were extensive differences within and between career ladders. Few universal causes of dissatisfaction were identified; the distinguishing facets were essentially unique to each

career ladder. For example, in the Intelligence Operations ladder, it was found that over half of the tasks airmen were trained to perform in mandatory technical training schools were never performed in the field. The study resulted in curriculum changes. Other training programs were reengineered based on the attitude and satisfaction studies, including those for the Automatic Tracking Radar Repairman ladder and the Disbursement Accounting ladder. In the Aircraft Control and Warning specialty, significant dissatisfaction among incumbents was linked to lack of specific task performance experience on the part of supervisors. The finding led to changes in previous job consolidation and merger decisions. In these specialties and others, dramatic improvements in job satisfaction resulted from remedial recommendations from job and attitude surveys.

The research project led to a methodology for identifying specialties with the greatest potential for job performance and reenlistment rate improvements through in-depth study of job attitudes and satisfaction. Specialty-unique profiles were designed based on the relationship between TAFMS (total active federal military service) and job interest, holding aptitude constant. The representative profile was for job attitudes to decrease with increasing tenure for first-term airmen, but then to increase with tenure for career airmen (see figure). The methodology allowed the identification of specialties with larger than typical "impact gaps" or with other kinds of relationships which strongly suggested that job satisfaction data would be useful in determining remedial interventions for improving job interests as the reenlistment decision point approached.
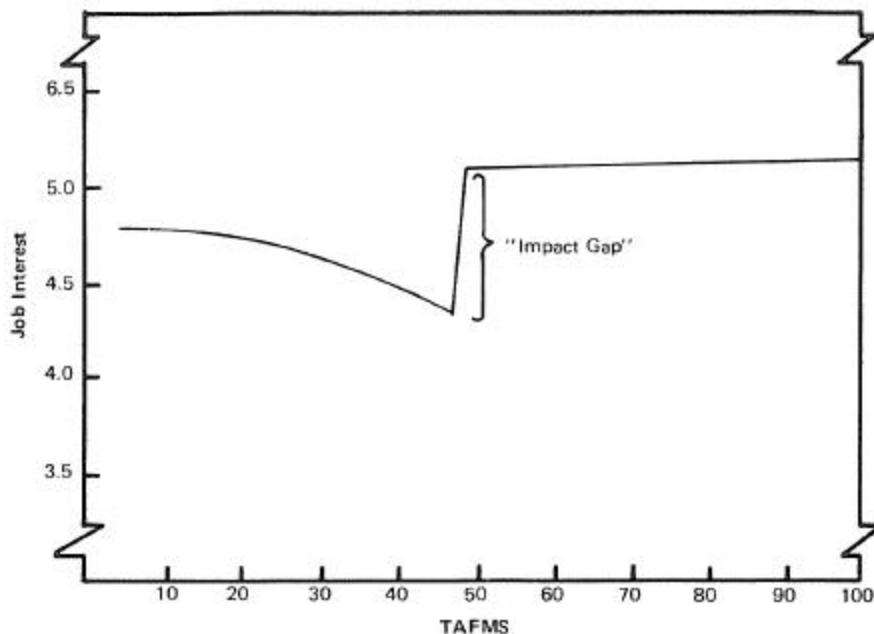
Another major finding concerned the relationship between statements of career intent/job attitudes and career decisions. About 53,000 first-term airmen respondents who had been surveyed at various years of service were tracked and their actual reenlistment decisions were determined. A comparison of reported intent to reenlist with actual "in/out" decisions reflected a significant relationship, with a large percentage of those saying "yes" staying and those saying "no" leaving. Career intent statements became more accurate over time; those obtained in the last two years of enlistment were more accurate than statements obtained during the first two years. In terms of job satisfaction research, the important finding was that career intent statements were sufficiently valid to permit their use as criterion for measuring effects of job reengineering actions. Further analyses showed the predictive value and accuracy of job attitude statements for anticipating surges and ebbs in reenlistments by occupational specialty.

Several additional important findings emerged from the job satisfaction research project. Seventy-five percent of airmen did not receive their stated "top three" pre-enlistment preferences as job assignments. Only 15 percent received their first preference. Airmen who received their first preference later reported being significantly more interested in their jobs and felt their talents and training were better utilized. The then-current job assignment system assigned few airmen to preferred work areas even though those who received preferred assignments tended to be more satisfied. Causal relationships between job attitudes and performance were established for selected specialties. In addressing the utilization of minorities, no practical within-specialty racial differences were found in types of job assignments and subsequent job attitudes.

Gould, R.B. (1976). *Review of an Air Force job satisfaction research project: Status report through September 1976* (AFHRL-TR-76-75). Lackland AFB, TX: Air Force Human Resources Laboratory. **E-16**

Tuttle, T.C., Gould, R.B., & Hazel, J.T. (1976). *Dimensions of job satisfaction: Initial development of the Air Force Occupational Attitude Inventory* (AFHRL-TR-75-1). Lackland AFB, TX: Air Force Human Resources Laboratory. **E-18**

II. RESEARCH ON AIR FORCE OFFICER PERSONNEL SYSTEMS

Precursors of the Air Force Officer Qualifying Test (AFOQT)

Aviation Psychology Program – World War II

In World War II, selection and classification of aircrew personnel became a pressing need. Before the war, qualification for pilot training was based on age, educational qualification (2 years of college), and a medical examination. The demand for pilots was low - less than 300 per year - and most of the work to select pilots was done by flight surgeons at the Army Air Corps School of Aviation Medicine. As world tension mounted and aircrew personnel requirements grew into the thousands, the Medical Division recommended the activation of a Psychological Research Agency to develop and validate new instruments for selecting pilots.

The Aviation Psychology Program was approved in June 1941 and developed during the next two years, first at Maxwell Field, Alabama, and later at additional sites. The most prominent psychologists and measurement specialists in the nation arrived from universities and testing agencies to lead the program. Many were given direct commissions at the rank of Major. Support personnel designated to work in the research centers were brought in from the officer and enlisted ranks of the Army Air Corps.[2]

The first products of the Aviation Psychology Program were initial screening tests for pilots, navigators, and bombardiers from the officer ranks in 1942 and in 1944 for gunners from the noncommissioned ranks. The selection tests were general intelligence tests and were given the names Aviation Cadet Qualifying Examination (ACQE) and Army Air Force Qualifying Examination (AAFQE). The purpose of the screening tests was to determine likelihood of success in flying training of young men with less than 2 years of college. Replacement of the previous 2-years of college requirement with scores attained on a general abilities test greatly expanded the applicant pool.

Once the men were selected for aircrew training, it was necessary to assign them to specific training courses. The Aircrew Classification Battery was developed for this purpose. The first classification battery was used in February 1942 and consisted of power and speeded paper-and-pencil tests as well as psychomotor tests (apparatus tests). In all, ten Aircrew Classification Batteries were used during World War II, each representing a modification of the preceding battery as additional empirical data accumulated. By 1944 separate classification composites had been developed for bomber and fighter pilots and for aerial gunners, air mechanic-gunners and radio operator-gunners. The last revision of the test battery during wartime was in June 1945. After V-J Day on 15 August 1945, input into pilot training ceased for six weeks.

After World War II, the research program was curtailed, and the staff turned its attention to documenting the wartime research. A 19-volume set of now classic books, called

---

[2] The mission of this program continued after the war and in the Air Force was accomplished by the Air Force Human Resources Laboratory and its predecessor and successor organizations.

Army Air Forces Aviation Psychology Program Research Reports, was completed. This series is often referred to as the "blue books," simply because the volumes were bound with blue covers. The program is considered a major milestone of applied psychology. Descriptions of test designs and completed studies from the Aviation Psychology Program appear in today's college textbooks on tests and measurement.

The aptitude and psychomotor tests developed in this era provided the foundation for modern aircrew selection tests used by the Air Force.

Rogers, D.L., Roach, B.W., & Short, L.W. (1986). *Mental ability testing in the selection of Air Force Officers: A brief historical overview*. (AFHRL-TP-86-23). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **O-01**

Valentine, L.D., Jr., & Creager, J.A. (1961). *Officer selection and classification tests: Their development and use*. (ASD-TN-61-145). Lackland AFB, TX: Personnel Laboratory, Aeronautical Systems Division. **O-02**

Post-War Officer Testing

After the war, research continued on two major lines of test development. The two lines were not independent; each branched and crossed with the other with respect to test content and use. One line consisted of the series of aircrew selection and classification devices started during World War II. The other line of development began in 1949 with the Aviation-Cadet Officer-Candidate Qualifying Test series. Initially, these tests were used for aircrew prescreening. Later, their use was expanded to include non-aircrew officer selection and classification. The two lines of test development eventually merged into a single test battery called the Air Force Officer Qualifying Test (AFOQT).

Activity on the first line of test development was focused on the Aircrew Classification Battery (ACB). Its use was resumed in October 1945 after the war ended and continued until 1947. From then until 1951 the two years of college requirement was reinstated for aircrew decisions. In 1951 operational testing on the ACB was resumed. The last aircrew battery continued until 1955. At that time psychomotor testing was discontinued and aircrew classification was based on the recently evolved AFOQT. Psychomotor testing was stopped due to problems keeping the equipment calibrated in mobile testing units.

The second line of test development after the war began in 1949 with a requirement for the Aviation-Cadet Officer-Candidate Qualifying Test (AC-OC-QT). It was intended for use in prescreening aviation cadet applicants, and the first two booklets were named the Aviation Cadet Qualifying Test (ACQT). About the same time needs developed for selecting officers for the Reserve Officer Training Corps and the Officer Candidate School.

In 1951 the first explicit use of the name Air Force Officer Qualifying Test in designating a set of test booklets occurred. The AC-OC-QT was incorporated and consisted of Officer Aptitude, Biographical Information, and Flying Aptitude test booklets. This preliminary version of the AFOQT was designed to partially fulfill the functions of an aircrew battery and to yield scores predictive of success in Officer Candidate School and in non-aircrew officer technical courses.

Rogers, D.L., Roach, B.W., & Short, L.W. (1986). *Mental ability testing in the selection of Air Force Officers: A brief historical overview*. (AFHRL-TP-86-23). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **O-01**

Valentine, L.D., Jr., & Creager, J.A. (1961). *Officer selection and classification tests: Their development and use*. (ASD-TN-61-145). Lackland AFB, TX: Personnel Laboratory, Aeronautical Systems Division. **O-02**

Air Force Officer Qualifying Test (AFOQT)

Chronology of Forms Developed (1951-2004)

Eighteen versions of the AFOQT were published by the Air Force Human Resources Laboratory (AFHRL) from 1951 until 1999 when the laboratory was closed. The chronology is summarized in the table with information about the test purpose, significant features and changes. Form Q was the last version published by AFHRL. All forms were administered as paper-and-pencil tests with separate answer sheets.

The practice was to document the development and standardization of each form in a separate technical report. The reports described test specifications, rationale for changes in test characteristics and procedures over previous forms, subtest content, number of items, and composite structure. Other topics were item writing, selection, and scoring, and statistical data on item difficulty, item discrimination, subtest reliability, composite reliability and test intercorrelations. Test standardization procedures, description of normative groups, and provisional and final conversion tables were also documented. These reports, along with additional empirical information gathered on the form while in operational use, served to guide development of the next successive form.

Year Implemented	AFOQT Form	Principal Use and Significant Features
1951	Preliminary Version	Used for aircrew classification and Officer Candidate School (OCS) and non-rated selection. Incorporated the Aviation-Cadet Officer Candidate- Qualifying Test (AC-OC-QT).
1953	A	Selection test for advanced AFROTC training (pilot, navigator, technical specialty). Had four interest scores (Administrative, Flying, Technical, and Quantitative).
1954	B	Selection test for first class of Air Force Academy (AFA), OCS, and advanced AFROTC training. Replaced Aircrew Classification Battery for selection of aviation cadets for pilot or observer training.
1956	C	Selection test for AFA, AFROTC, OCS and direct appointment, aircrew, and Air National Guard (ANG) and Air Reserve.
1957	D	Selection test for AFA, AFROTC, OCS and direct appointment, aircrew, and Air National Guard (ANG) and Air Reserve.
1958	E	Selection test for AFA, AFROTC, OCS and direct appointment, aircrew, Air National Guard (ANG) and Air Reserve, as well as the Air Force's new Officer Training School (OTS) program.
1959	F	Selection test for AFA, AFROTC, direct appointment, aircrew, Air National Guard (ANG) and Air Reserve, and OTS. Observer-Technical composite was renamed Navigator-Technical.
1960	G	Selection test for AFA classes graduating in 1959 through 1960. Then AFOQT replaced by College Entrance Examination Board

		(CEEB) test for AFA selection. Continued AFOQT use for selecting AFROTC, direct appointment, aircrew, Air National Guard (ANG) and Air Reserve, and OTS.
1964	AFOQT-64	Selection test for OTS and AFROTC; classification test for pilot and navigator training
1966	AFOQT-66	Selection test for OTS and AFROTC; classification test for pilot and navigator training. New norm group based on Project TALENT battery.
1968	AFOQT-68	Selection test for OTS and AFROTC; classification test for pilot and navigator training. Added three sets of conversion tables for educational level norms. Test manual published on interpretation and utilization of scores on the AFOQT.
1970	K	Selection test for OTS and AFROTC; classification test for pilot and navigator training.
1972	L	Selection test for OTS and AFROTC; classification test for pilot and navigator training.
1975	M	Selection test for OTS and AFROTC; classification test for pilot and navigator training.
1978	N	Selection test for OTS and AFROTC; classification test for pilot and navigator training. New normative group for AFOQT. Officer Biographical Inventory removed. Removed one set of educational level conversion tables; retained two levels of educational norms.
1981	O	Selection test for OTS and AFROTC; classification test for pilot and navigator training. Pilot Biographical Inventory and educational level conversion tables removed. Officer Quality composite renamed Academic Aptitude.
1987	P (P1 and P2)	Selection test for OTS and AFROTC; classification test for pilot and navigator training. Two parallel versions, Information Pamphlet for Examinees, and test manual were published.
1994	Q (Q1 and Q2)	Selection test for OTS and AFROTC; classification test for pilot and navigator training. Last form published by AFHRL.

Berger, F.R., Gupta, W.B., Berger, R.M., & Skinner, J. (1990, April). *Air Force Officer Qualifying Test (AFOQT) Form P: Test Manual* (AFHRL-TR-89-56, AD-A221 004). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-03**

Glomb, T.M., & Earles, J.A. (1997). *Air Force Officer Qualifying Test (AFOQT) Forms Q: Development, Preliminary Equating and Operational Equating* (AL/HR-TP-1996-0036). Brooks AFB, TX: Armstrong Laboratory. **O-04**

Rogers, D.L., Roach, B.W., & Short, L.W. (1986). *Mental ability testing in the selection of Air Force Officers: A brief historical overview* (AFHRL-TP-86-23). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-01**

Valentine, L.D., Jr., & Creager, J.A. (1961). *Officer selection and classification tests: Their development and use*. (ASD-TN-61-145). Lackland AFB, TX: Personnel Laboratory. **O-02**

Subtests, Scoring, and Composites

The subtest content of AFOQT forms varied over the years but typically consisted of tests of verbal, quantitative, pilot and navigator aptitude. Among the tests of verbal abilities were reading comprehension, verbal analogies, vocabulary, and English usage. Quantitative aptitude was tested with general mathematics, interpretation of data, and arithmetic reasoning. Tests of pilot abilities included aviation information, mechanical principles, visualization of maneuvers, instrument comprehension, flight orientation, aerial landmarks, stick and rudder orientation, and table and scale reading. Navigator abilities were measured with quantitative tests, data interpretation, general science, mechanical principles, aerial and spatial orientation, and scale reading. Both power and speeded subtests have been used to assess verbal, quantitative, and aircrew aptitudes.

Separate biographical inventories for officers and pilots were included in all forms of the AFOQT through Form M. These inventories were composed of activities associated with males, normed on male only samples, and taken only by male examinees. Unable to remove the sex bias from the items, the Officer Biographical Inventory was removed from Form N. The Pilot Biographical Inventory was retained in Form N but was dropped from Form O because of low validities and probable sex and racial bias of the subtest. Decisions about removing the biographical inventories were prompted by increasing numbers of women entering military service.

The method used to score subtests was "rights only" for biographical inventories. In most of the early forms a correction for guessing formula was used with all other subtests, both power and speeded. Later forms tended to use the correction for guessing scoring method only on subtests specifically designated as speeded subtests. However, beginning with Form O, all subtests were scored "rights only," because no subtests were judged to be purely speeded.

Five composite scores have been obtained from the AFOQT since Form A was produced: Pilot, Navigator-Technical, Academic Aptitude, Verbal and Quantitative. In Forms A through G, the Navigator-Technical composite was called Observer-Technical. The Academic Aptitude composite combines Verbal and Quantitative scores. Until Form O when it was renamed to prevent misinterpretation of what the composite was intended to measure, the Academic Aptitude composite was called the Officer Quality composite.

Composite scores for Forms O, P, and Q were reported on a percentile scale (1-99). Prior forms used a percentile scale with scores reported in 5-point increments (1, 5, 10,…, 95). The earliest forms of the AFOQT used a stanine scale. Scores on all five composites have been derived for all applicants since Form O was implemented. This was made possible by reducing the number of items in Form O to 380 from the 606 used in Form N, thereby decreasing testing time, as well as by printing the entire Form O test in a single booklet. Prior forms split the content into separate booklets, usually five.

The subtests and composites of the AFOQT (Forms O, P, and Q) in use from 1981 through 2004 are shown in the table.

Subtest	No. of Items	Composites				
		Pilot	Navigator-Technical	Academic Aptitude	Verbal	Quantitative
Verbal Analogies	25	X		X	X	
Arithmetic Reasoning	25		X	X		X
Reading Comprehension	25			X	X	
Data Interpretation	25		X	X		X
Word Knowledge	25			X	X	
Math Knowledge	25		X	X		X
Mechanical Comprehension	20	X	X			
Electrical Maze	20	X	X			
Scale Reading	40	X	X			
Instrument Comprehension	20	X				
Block Counting	20	X	X			
Table Reading	40	X	X			
Aviation Information	20	X				
Rotated Blocks	15		X			
General Science	20		X			
Hidden Figures	15		X			
Total	380					

Glomb, T.M., & Earles, J.A. (1997). *Air Force Officer Qualifying Test (AFOQT) Forms Q: Development, Preliminary Equating and Operational Equating* (AL/HR-TP-1996-0036). Brooks AFB, TX: Armstrong Laboratory. **O-04**

Rogers, D.L., Roach, B.W., & Short, L.W. (1986). *Mental ability testing in the selection of Air Force Officers: A brief historical overview* (AFHRL-TP-86-23). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-01**

Valentine, L.D., Jr., & Creager, J.A. (1961). *Officer selection and classification tests: Their development and use*. (ASD-TN-61-145). Lackland AFB, TX: Personnel Laboratory. **O-03**

Norms and Standardization (Form A through Form Q)

Several normative bases have been used with the AFOQT. In 1955, the AFOQT was normed using cadets at the newly established Air Force Academy (AFA) as the reference group. The AFA cadets were used until 1960 when the requirement for the AFOQT as a selection test was eliminated in favor of the College Entrance Examination Board (CEEB), an early name for the college admission test battery in the Scholastic Aptitude Test program.

In anticipation of the loss of the AFA as a reference group, a new normative base was obtained by administering Form G of the AFOQT and the Project TALENT test battery to more than 5,000 applicants for the AFA class entering in 1960. The Project TALENT tests were ability and aptitude tests used in a national survey of about 400,000 students of high school age. A subsequent indirect method of using TALENT composites and basic airmen samples became the accepted procedure for standardizing successive forms of the AFOQT.

When AFOQT Form N was developed with substantial content changes, a new standardization sample was necessary. It was composed of basic airmen, AFROTC cadets, OTS cadets, AFA cadets, and junior officers. The sample was designed to represent the full range of ability expected in the officer applicant population and included subjects from all major sources for Air Force commissioning and specialized training programs.

Composite scores on AFOQT Forms O, P, and Q were linked to Form N scores and the normative group using equipercentile equating methods. A common item or anchor item design was used for Form O and an equivalent groups design was used for Form P and later Form Q by administering the forms in the same testing sessions with AFOQT tests that had been previously equated to Form N. In the case of Form P the equating was through Form O and in the case of Form Q, the linkage was through Form P1.

A new normative group was established for AFOQT Forms R and S which were developed for the Air Force under contract by the Operational Technologies Corporation. The change was necessary to update norms for the revised content and structure introduced in the officer testing program in 2005.

Gould, R.B. (1978). *Air Force Officer Qualifying Test Form N: Development and standardization* (AFHRL-TR-78-43, AD-A059 746). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-05**

Rogers, D.L., Roach, B.W., & Short, L.W. (1986). *Mental ability testing in the selection of Air Force Officers: A brief historical overview*. (AFHRL-TP-86-23). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **O-01**

AFOQT Form S – The Current Test

Form S of the AFOQT has been in the field for officer selection and classification testing since 2005. Development of the form was completed under contract. The project goal initially was to develop Form R as a replacement for Form Q with comparable content, composite structure, and testing schedule. Work proceeded for several years toward that goal. As test development neared completion, the plan to introduce Form R as essentially parallel to Form Q was changed. The Air Force directed substantive improvements which resulted in implementation of Form S with (1) reduced test content to shorten testing time, (2) revised selection composite structure, (3) refined scoring procedures for the aircrew classification composites, (4) updated reference group (normative base), and (5) experimental non-cognitive content. Major features of Form S are summarized in the table on the next page.

Form S has cognitive subtests distributed across five selection and classification composites: Verbal, Quantitative, Academic Aptitude, Pilot, and Navigator-Technical. The subtests contain anchor items drawn from earlier forms of the AFOQT as well as items which were newly written to ensure currency of subject matter tested and comprehensive coverage of cognitive domains important for officer performance. The 11 cognitive subtests in Form S are a subset of the 16 which appeared in previous forms of the AFOQT. Analyses showed that the same factor structure and comparable reliability of the composites were achievable using the reduced set of tests. The benefit of the streamlined test booklet is substantial shortening of administration time.

Follow-on analyses were completed to determine if validities for the aircrew classification composites (Pilot and Navigator-Technical) could be improved with the reduced battery and with alternate scoring procedures. Training performance and completion criteria were obtained for samples of rated officers attending undergraduate pilot or navigator training. Findings were that predictive effectiveness could be increased by reconfiguring the subtest structure of the aircrew classification composites and by replacing unit weights with regression-based subtest weights for composite scoring. The content of the composites was revised to place greater emphasis on quantitative skills in the Pilot composite and on verbal skills in the Navigator-Technical composite. Two subtests (Rotated Blocks and Hidden Figures) formerly scored in the composites were retained in the test booklet as experimental measures.

The norm sample data for earlier AFOQTs (Form N through Q), which were collected in the 1970s, were replaced with "new millennium" norm group data in preparation for implementing Form S. Selected to be representative of the ability range of applicants for officer commissioning and flying training programs, the updated norm group consists of nearly 2,500 military examinees administered the newly configured battery of tests for the AFOQT. The composite percentile scores, which are currently being used by officer selection boards, are based on the new norm group.

A final feature of the current AFOQT is the inclusion of a non-cognitive test called the Self-Description Inventory (SDI+). The SDI+ is an experimental adjunct to the battery

and is designed to assess major personality dimensions. Future studies will explore the utility of the traits for the officer testing program.

The data, which are presently accumulating from officer applicants being tested on AFOQT Form S, will support an essential follow up evaluation of how well the test is operating for Air Force officer selection and classification. Critical research questions about the reliability, validity, and equity of the battery need to be addressed to ensure the properties of Form S comply with national testing standards.

AFOQT Form S Testing Schedule and Structure

				AFOQT Composites				
Subtest	Test/Activity Description	Number of Items	Testing Time in Minutes [a]	Pilot	Nav-Tech	Academic	Verbal	Quantitative
	Pre-Test Activities		24					
1	VA – Verbal Analogies	25	9		X	X	X	
2	AR – Arithmetic Reasoning	25	30	X	X	X		X
3	WK – Word Knowledge	25	6			X	X	
4	MK – Math Knowledge	25	23	X	X	X		X
5	IC – Instrument Comprehension	20	9	X				
6	BC – Block Counting	20	5		X			
7	TR – Table Reading	40	9	X	X			
8	AI – Aviation Information	20	9	X				
9	GS – General Science	20	11		X			
10	RB – Rotated Blocks	15	15					
11	HF – Hidden Figures	15	10					
12	SDI – Self-Description Inventory	220	40					
	Collection of Materials & Break		12					
TOTAL		470	213					
	Testing Time: 3 hr 33 mins							

[a] Subtest times listed include subtest directions and test performance.

Alley, W.E. (2002). *Development of experimental measures for the AFOQT.* San Antonio, TX: Operational Technologies Corporation. **O-06**

Shore, C.W., & Gould, R.B. (2003). *Developing pilot and navigator/technical composites for the Air Force Officer Qualifying Test (AFOQT) Form S.*   San Antonio, TX: Operational Technologies Corporation. **O-07**

Shore, C.W., & Gould, R.B. (2002). *Reduction of Air Force Officer Qualifying Test (AFOQT) administration time*. San Antonio, TX: Operational Technologies Corporation. **O-08**

Weissmuller, J.J., Schwartz, K.L., Kenney, S.D., Shore, C.W., & Gould, R.B. (2004). Recent developments in USAF officer testing and selection. *Proceedings of 46th Annual Conference of the International Military Testing Association and NATO's Research Task Group on Recruiting and Retention of Military Personnel*, p. 268-279. Brussels, Belgium. **O-09**

Minimum Qualifying Scores

Establishing selection standards and setting minimum qualifying scores for officer programs are policy matters. The AFHRL research report series contains relatively little information on the topic and is not useful for either tracing the origin of specific qualifying scores or changes in standards.

In a test manual for the AFOQT published in 1969, the author noted that for Air Force tests, minimum qualifying scores were established by Headquarters, United States Air Force, and were promulgated by directive. At that time, qualifying scores were set on one or more composites in nearly all selection and classification programs for which the AFOQT was used. Exceptions were the Verbal and Quantitative composites which had no minimum qualifying scores for any program. Miller (1969, p. 27) stated:

> "Minimum qualifying scores are not the same in all programs, and they are subject to change at any time. Changes are made in accordance with the availability of applicants for the various programs and the needs of the Air Force. When there are many applicants to fill a small quota, minimum qualifying scores may be set very high. If the need for personnel to fill a quota is such that most applicants must be accepted, minimum qualifying scores must be set very low. In this case applicants with mediocre or borderline aptitudes are entered into the program, and it can be expected that the elimination rate will rise."

After publication of the 1969 report, qualifying scores were added for the Verbal and Quantitative composites of the AFOQT. Currently, the minimum qualifying scores used to select applicants for pre-commissioning training programs are the 15th percentile on the Verbal composite and 10th percentile on the Quantitative composite. The cutoffs are set at low values on these composites to permit flexibility in meeting Air Force recruiting objectives. AFOQT scores are also one of several factors considered in evaluating candidates for rated training. The AFOQT composite scores needed to qualify for acceptance for undergraduate pilot training are a minimum Pilot composite score at the 25th percentile and a minimum Navigator-Technical composite score at the 10th percentile. Also, the applicant's combined Pilot and Navigator-Technical scores must be at least 50. To qualify for undergraduate navigator training, the applicant must achieve a minimum Navigator-Technical score corresponding to the 25th percentile, a minimum Pilot composite score at the 10th percentile, and a combined Navigator-Technical and Pilot score of 50. Based on the personal recollections of senior managers formerly on the AFHRL research staff, the Air Force has used these standards for about 30 years. It should also be noted that the Air Force has set minimum qualifying scores on selection factors besides the AFOQT.

Miller, R.E. (1969). *Interpretation and utilization of scores on the Air Force Officer Qualifying Test* (AFHRL-TR-69-103, AD 691 001). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **O-10**

AFOQT Research Topics

Educational-Level Norms

Educational-level norms were introduced in the AFOQT testing program with AFOQT-68 and continued to be used in the next four forms – AFOQT Forms K, L, M, and N. Initially, separate conversion tables were used for three educational level groups of officer applicants: examines with less than 2 years of college; examinees with 2 or more years of college but not college graduates; and examinees who were college graduates. Later, the number of conversion tables was reduced to two in AFOQT Form N, one set for examinees with less than 2 years of college and one set for examinees with 2 or more years of college including college graduates. Educational-level norms were dropped with the implementation of AFOQT Form O in 1980. Successive forms of the AFOQT (Form P, Q, R, and S) were also implemented without educational-level conversion tables.

Changes in the use of separate scoring tables for educational levels were based on studies conducted from 1968 to 1986 which yielded conflicting results about the effect of educational level. Although differences in testing populations and experimental designs were noted between the studies, each study may have provided accurate results for the time and test form of interest. However, the collective implications for the current operational test, AFOQT Form S, are not clear. Additional research is needed to determine if educational level norms would be appropriate for use with the operational test or with the next revision (Form T). The required analyses should be conducted as part of a development plan for Form T and need to address whether relationships between test scores and performance criteria are the same for different educational levels.

Skinner, J. (2007). *Educational level norms.* San Antonio, TX: Operational Technologies Corporation. **O-11**

Predictive Validity

Composite scores from the Air Force Officer Qualifying Test (AFOQT) are used for selection to pre-commissioning training in Officer Training School (OTS) and Air Force Reserve Office Training Corps (AFROTC) and to follow-on rated training in Undergraduate Pilot Training (UPT) and Undergraduate Navigator Training (UNT). Inferences made from the test scores are that officer applicants with higher test scores are more likely to successfully complete military training programs. National standards for the use of personnel selection procedures recommend criterion-related validity studies as evidence to support such inferences.

Numerous studies have been conducted on the AFOQT, and the results have consistently shown that the composites are correlated with officer training performance measures. In the OTS program, the Academic Aptitude, Verbal, and Quantitative composites have been found to predict graduation/elimination, training effectiveness ratings from instructors, academic grades, and whether cadets were distinguished graduates. Test scores for AFROTC cadets have also been shown to correlate with training completion/non-completion, an instructor rating of training performance, and distinguished graduate status in the Professional Officer Course. In studies of pilot training, students with higher AFOQT Pilot composite scores had higher probabilities of completing training, achieved higher class ranks, and required fewer flying hours to achieve proficiency. The findings for the Navigator-Technical composite in UNT were similar. Validities were significant for several performance criteria including training outcome (graduation/elimination), average classroom lesson score, average simulator lesson score, and average flying lesson score. Course grades in non-rated technical training were also well predicted by AFOQT composites.

Collectively, the study results demonstrate the validity of the AFOQT as an Air Force officer selection and classification instrument. Air Force managers, who are expected to continue to use AFOQT test scores to set entry standards and inform applicant selection and classification decisions, can point to the body of evidence from prior validity studies to defend the value and use of the AFOQT. To insure continued compliance with national testing standards, it will be necessary to complete updated studies of the criterion-related validity of the test when new forms are published.

Skinner, J. (2006). *Criterion-related validity of the operational AFOQT composites*. San Antonio, TX: Operational Technologies Corporation. **O-12**

Retesting

Air Force policy allows retests on the Air Force Officer Qualifying Test (AFOQT) after 180 days have passed from the date the test was previously administered (AFI 36-2605, 14 November 2003). An alternate version of the battery is given whenever possible. Only two test administrations are authorized, but waivers may be granted. The requests must contain details showing that the applicant's potential abilities in subjects relevant to the AFOQT have changed substantially since the last test administration; for example, by attending college courses or acquiring flying experience.

The Air Force maintains a data base containing records of all AFOQT administrations. The data support research on amount of retesting, intervals between retests, score changes for retesters, test-retest reliability, and predictive validity estimates.

From 1981 to 1995, about 16 percent of AFOQT examinees were retesters. The breakdown for about 280,000 test administrations was 84% tested once (non-retesters), 13% tested two times, 2% tested three times, and 0.4% tested four or more times.

Requests to retake the AFOQT were self-initiated, and presumably examinees were trying to increase their scores and improve their chances of being selected for training programs including specialized rated programs. Studies showed that retesters' initial scores were significantly lower than those of non-retesters. The first score versus subsequent administration scores indicated a clear pattern of score increases on all composites. The largest improvements were on the Pilot and Navigator-Technical composites. Scores increased most for examinees with lower aptitudes. The most substantial subtest gains were in Instrument Comprehension, a subtest in the Pilot composite. Arithmetic Reasoning, Math Knowledge, and General Science were among the subtests with the smallest gains. Appreciable improvements in Instrument Comprehension as well as in Aviation Information, both of which test pilot job knowledge, may reflect efforts on the part of retesters to learn specialized material between testing sessions.

The studies showed that individuals seeking opportunities to retest on the AFOQT were different from non-retesters. Retesters tended to have lower scores initially. Although their scores improved with subsequent retests, their retest scores often remained lower than non-retesters.

Retest data were used to evaluate consistency of test measurements. The test-retest reliability reflects the stability of a measure administered more than once and is estimated by correlating the scores obtained from a group of examinees administered the same test on two occasions. Test-retest correlations were estimated to be Pilot = .83, Navigator-Technical = .87, Academic Aptitude = .89, Verbal = .89, and Quantitative = .84. Overall, the reported reliabilities are moderately high and suggest stability in AFOQT measurements across time. Carry-over effects due to motivation are probable. The retesters were a self-selected group who requested an opportunity to retake the test.

The predictive effectiveness of first, last, and averaged AFOQT scores against a final outcome criterion (pass/fail) in Undergraduate Pilot Training was evaluated. Correlation coefficients for the Pilot and Navigator composites were statistically significant for the three conditions. The highest validities obtained were for average tests scores. For retesters the average of all scores achieved on the Pilot and Navigator-Technical composites provided a more valid index of potential training success than the last time tested score seen in practice by Air Force pilot selection boards.

The designs used in the retest studies did not permit the effects of motivation, learning, and maturation to be addressed. Follow-on studies are recommended which use random assignment of subjects to retest groups, rather than self-selected examinees. Further, more accurate estimates of AFOQT test-retest reliability could be obtained.

Skinner, J. (2006). *AFOQT retesting*. San Antonio, TX: Operational Technologies Corporation. **O-13**

Test Bias

The basic assumption of the classical model of selection is that scores on employment tests are linearly related to measures of performance in the employment setting. Job applicants with higher scores on the test are expected to perform at higher levels in training courses and on the job. Predictive validity studies of the AFOQT found positive relationships between the selection composites of the AFOQT and officer performance criteria. Other studies examined whether subgroup differences existed for Air Force applicants in the aptitude factor structure of the AFOQT or in scores obtained on the test. Comparisons of aptitude factor structure revealed near identity of the structure of abilities measured by the AFOQT for gender and ethnic groups. Nevertheless, mean test score differences for minority and majority subgroups were found.

Combining the issues of the predictive validity of the test and subgroup differences addresses the separate question of whether the AFOQT is equitably predictive for majority and minority subgroups, regardless of mean test score differences. The accepted model for evaluating test bias is called Cleary's regression model. If the relationship or regression of tests scores against the performance criterion the test is designed to predict is the same for majority and minority subgroups, the test is found to be nonbiased or equitable. If the relationship is not the same, that is, it varies between subgroups, the presence of bias is detected.

Three studies of the AFOQT were completed which specifically addressed the question of test bias. Criteria were performance measures in Officer Training School, Undergraduate Pilot Training, and Undergraduate Navigator Training. Results of the AFOQT studies were consistent with those from the literature on standardized tests. The overall conclusion was that the officer test was equitable for males and females and for majority and minority examinees. For criterion/test relationships where there was evidence of test bias, it was predominately in the form of level differences with the overprediction of minority subgroup performance. Differences were not appreciable, and discrimination against minority subgroups was not supported by the data. Overall, the studies indicated that continued use of the AFOQT in a consistent manner by Air Force officials would result in fair selection.

Test bias studies should be continued on new AFOQT forms as criterion data mature on sufficient numbers of officers. Questions about subgroup performance on the test will continue to be asked as the composition of the officer force becomes more diverse. An ongoing research program on the AFOQT is needed to provide the answers.

Alley, W.E. (2006). *Conducting a test bias study*. San Antonio, TX: Operational Technologies Corporation. **RM-09**

Skinner, J. (2006). *Test bias studies of the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-14**

Utility Analysis of Pilot and Navigator-Technical Composites

Researchers use different ways to explain and measure the value of the Air Force Officer Qualifying Test (AFOQT) for aircrew selection and classification. The purpose is to help customers and test users understand the benefit of a merit-based selection system with standardized test scores as the foundation over a random process of admitting candidates to training on a first come, first serve basis. Because national testing standards recommend significance testing of the validity coefficient to empirically demonstrate the relationship between a selection test like the AFOQT and the performance criterion it is designed to predict – success in aircrew training – correlation coefficients are routinely calculated. To facilitate interpretation, the correlation coefficients are often supplemented with tabular and graphic techniques to show the increased probability of completing training for candidates with higher AFOQT test scores. Expectancy tables, line graphs and bar charts are routinely used. Another approach is to put a "dollar value" on success or failure in training to demonstrate the utility of the AFOQT.

An early utility analysis estimated the number of examinees disqualified by the Pilot composite who would have been eliminated had they entered training in the late 1960s. The number was 365. At an estimated average cost per eliminee of $24,000, the total savings in one year from using the Pilot composite was $8,760,000.

In another cost avoidance study, the performance criterion was "extra flying hours" in Undergraduate Pilot Training (UPT). A significant negative correlation was found between Pilot Candidate Selection Method (PCSM) scores, which include the AFOQT Pilot composite, and extra hours flown to reach proficiency on navigation check rides in the T-37 and in the T-38. On average pilot trainees who scored in the top two quintiles on PCSM required no extra flying hours, while those in the bottom three quintiles did. Using costs provided by HQ AETC, the remedial sorties were calculated to cost about $1,000,000. The study showed that higher costs are incurred for lower ability candidates during training and demonstrated the financial benefits from using tests, in this case the PCSM, for personnel screening.

A broader approach focused on the monetary value of the increased productivity that can be realized by selecting and training better quality applicants. Utility formulae were used that considered the cost of testing, the cost savings due to decreased attrition in UPT, and the dollar value expected from increased productivity of the new pilots. The utility concepts and formulae were introduced in the 1940s, improved in the 1950s, and gained acceptance with wider application in the 1970s after breakthroughs were made in how to accurately estimate some of the components. The study showed the value of the AFOQT for pilot selection during FY82 to be more than 100 million dollars over the five-year period of obligation of the new pilots. A similar analysis was completed of the utility of the Navigator-Technical composite for selecting candidates for the Undergraduate Navigator Training (UNT) program. Although the estimated value was lower for navigators than for pilots, the dollar benefits were still substantial, in the range of $10 to $15 million over the 5-year period of obligation for navigators completing UNT during

the 1980s. The formulas demonstrated that after considering the cost of testing, a one-time expense, the benefits from selecting higher quality candidates with the aid of an ability test accrue not only from reduced attrition in training but also from higher productivity throughout the tenure of the candidates.

Utility analyses are an effective way to demonstrate the value of a selection test to senior managers in an organization. The main drawback is that cost figures from these types of studies are subject to rapid obsolescence.

Skinner, J. (2007). *Utility analysis of the AFOQT Pilot and Navigator Composites*. San Antonio, TX: Operational Technologies Corporation. **O-15**

Officer Screening Tests

Officer screening tests are no longer used by the Air Force, but they were an important part of the testing program before technology advancements allowed hand-scored test booklets and answer sheets to be replaced with more efficient and accurate automated scanning and computer scoring. Several screening or short-form versions of the Air Force Officer Qualifying Test (AFOQT) were developed for the officer testing program. The goal was to reduce time and costs associated with applicant testing and processing. Recruiters and examinees were provided with preliminary score results from the screening tests while waiting for official score reports. The benefit was to allow recruiters to eliminate processing delays for potentially qualified applicants.

A common feature in designing screening tests was accuracy to predict performance on composites derived from the full-length AFOQT battery. The earliest screening tests, which were developed in the 1960s, were separate tests containing unique items with similar content and difficulty as those in the corresponding subtests on the full-length AFOQT. There was no overlap in the items tested on the short-form and full-length tests. Later, in the 1980s, the procedure changed to developing the screening tests with overlapping content by identifying the items from the full-length battery. Screening tests for officers were eliminated in the 1990s when AFOQT computer-scoring was centralized.

In the event of a future mobilization, screening tests could be reinstituted to efficiently process large numbers of candidates by identifying those who should be disqualified, assigned immediately to certain jobs, or tested further for specialties with higher cognitive demands.

Skinner, J. (2006). *Officer screening tests.* San Antonio, TX: Operational Technologies Corporation. **O-16**

Rated Officers

Pilot Candidate Selection Method (PCSM)

The Pilot Candidate Selection Method (PCSM) is used by the Air Force to identify the best qualified pilot training applicants. The PCSM algorithm has paper-and-pencil aptitude test scores, computer-based cognitive and psychomotor scores, and a measure of previous flying experience. The measures are combined in a regression equation which ranks applicants on probable success in flying training. The algorithm was developed in the 1980s, refined in the 1990s, implemented in 1993, and updated in 2006. Paper-and-pencil testing measures are obtained from the Pilot composite of Air Force Officer Qualifying Test. The Basic Attributes Test (BAT) was the source of computer-based test scores from 1993 until 2006, when it was replaced by the Test of Basic Aviation Skills (TBAS). Several studies were completed to identify measures and weights for the algorithm and to demonstrate PCSM validity for pilot selection. Studies showed the AFOQT scores offered the most predictive effectiveness followed by flying experience and psychomotor skills.

Carretta, T.R., & Ree, M.J. (1993) *Pilot candidate selection method (PCSM): What makes it work?* (AL-TP-1992-0063). Brooks AFB, TX: Manpower and Personnel Research Division, Human Resources Directorate, Armstrong Laboratory. **O-17**

Basic Attributes Test (BAT)

In 1955, the Air Force discontinued apparatus-based testing for aircrew selection and classification due to administrative problems with equipment calibration essential for accurate assessment of applicants' abilities. Testing continued with the paper-and-pencil Air Force Officer Qualifying Test battery. Prior to that time and continuing back to World War II, measurement of psychomotor abilities with apparatus tests was an integral part of aircrew ability testing. With advances in computer technology in the 1970s, interest was renewed in psychomotor testing. Updated studies by the Air Force Human Resources Laboratory in the 1970s confirmed their utility for pilot selection. Further, as testing theory advanced, additional cognitive and psychological factors were identified that were believed to be related to flying training outcomes.

In 1981, a variety of experimental aircrew selection tests were assembled in a battery called the Basic Attributes Test (BAT) and prepared for computer administration to examinees. Prototype BAT stations at centralized testing locations were supplemented with portable testing units called Porta-BAT, which were easily transportable and allowed for decentralized testing. Prototypes of the BAT (and Porta-BAT) contained 15 tests measuring psychomotor skills, information processing abilities, and personality/attitude characteristics.

As studies accumulated on the structure and validity of the BAT, the number of tests identified for operational use was reduced to five. Three were psychomotor ability tests. The first was a rotary pursuit task measuring multi-limb coordination called the Two-Hand Coordination Test. The second psychomotor test, named Complex Coordination, measured control precision and multi-limb coordination by using right and left hand control sticks. The Time Sharing Test measured reaction time and rate control. Information processing ability was assessed with the Item Recognition Test, a short term memory test. Personality/attitude characteristics were measured with the Activities Interest Inventory, an indicator of attitude toward risk taking.

The BAT was used as part of the pilot selection system from 1993 to 2006 when it was replaced by the Test of Basic Aviation Skills (TBAS).

Carretta, T.R. (1987). *Basic Attributes Tests (BAT) system: Development of an automated test battery for pilot selection* (AFHRL-TR-87-9, ADA185649). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **O-23**

Carretta, T.R., & Ree, M.J. (1993) *Pilot candidate selection method (PCSM): What makes it work?* (AL-TP-1992-0063). Brooks AFB, TX: Manpower and Personnel Research Division, Human Resources Directorate, Armstrong Laboratory. **O-17**

Test of Basic Aviation Skills (TBAS)

The Test of Basic Aviation Skills (TBAS) is a computer-based test of flight aptitude developed by HQ AETC/SAS and implemented for selection of pilot trainees in 2006. The TBAS consists of nine subtests measuring psychomotor (hand/eye coordination), cognitive (spatial ability), short-term memory, and multi-task performance. Before implementing TBAS, several research studies were completed to insure that it was a suitable replacement test for the Basic Attributes Test (BAT). Beginning in 1993, the BAT was used in the Pilot Candidate Selection Method (PCSM). Formerly, the PCSM combined scores from the Air Force Officer Qualifying Test (AFOQT) Pilot composite with scores from the BAT and a measure of prior flight experience in a regression-weighted pilot aptitude composite. Studies of the TBAS addressed validity, reliability, and updating the weights in PCSM. Results supported use of the TBAS in the PCSM to keep test content current for pilot selection, to avoid compromise that inevitably occurs as a consequence of leaving a test like the BAT in the field for a long period of time, and to take advantage of improvements in computer hardware and software for computer-based test administration. The TBAS is in the early stages of operational use, and additional research requirements exist. These include documentation of the test development program, analyses of test-retest reliability, development of norms, and examination of gender and ethnic subgroup performance. Long range research plans are to develop TBAS II with expanded test content and explore classification utility for pilot training track selection, as well as for other technical specialties (for example, navigator, air battle manager, air traffic controller).

Carretta, T.R. (2005). *Development and validation of the Test of Basic Aviation Skills* (AFRL-HE-WPTR-2005-0172, ADA442563). Wright-Patterson AFB, OH: Human Effectiveness Directorate, Air Force Research Laboratory. **O-18**

Ree, M.J. (2004a). *Making scores equivalent for Test of Basic Aviation Skills (TBAS) and Basic Attributes Test (BAT)*. San Antonio, TX: Operational Technologies Corporation. **O-19**

Ree, M.J. (2004b). *Reliability of the Test of Basic Aviation Skills* TBAS. San Antonio, TX: Operational Technologies Corporation. **O-20**

Ree, M.J. (2003a). *Test of Basic Aviation Skills (TBAS): Incremental validity beyond the Air Force Officer Qualifying Test (AFOQT) Pilot composite for predicting pilot criteria*. San Antonio, TX: Operational Technologies Corporation. **O-21**

Ree, M.J. (2003b). *Test of Basic Aviation Skills (TBAS): Scoring the tests and compliance of the tests with the standards of the American Psychological Association*. San Antonio, TX: Operational Technologies Corporation. **O-22**

Situational Awareness

Within the Air Force Human Resources Laboratory, there was a large research program related to learning abilities measurement that had an applied and operational focus on aircrew (pilot and navigator) selection. During the 1990s there began several collaborative attempts to incorporate test batteries developed in the Learning Abilities Measurement Program (LAMP) into a separate research unit on aircrew selection. These collaborations were beginning to show promise before the closing of the laboratory.

A major collaboration was a validation study to predict situational awareness in 171 F-15 pilots. This was one of the most important and high-visibility studies ever conducted not only with LAMP tests but also on aircrew selection. It also represented a unique opportunity to test F-15 pilots. The study was requested by the US Air Force's Chief of Staff, General Merrill McPeak, to investigate pilot situational awareness in combat. "Loss of situational awareness" was the most frequently cited reason for accidents and mishaps, and so the study was ordered to investigate how pilots might develop and maintain situational awareness. There were several kinds of investigations of the problem, but a major focus was on pilot selection. LAMP staff members joined up with several other research groups to put together a comprehensive information-processing test battery that could be administered to fighter pilots and validated against supervisor- and peer-ratings of situational awareness. Flight experience was statistically held constant.

A total 171 F-15 fighter pilots were administered a comprehensive 5-hour battery over two days at their duty locations. Hundreds of basic trainees were also administered the battery, in parts. The battery consisted of 18 cognitive and 5 psychomotor tasks (most of which were developed in the LAMP project), along with other LAMP tests designed specifically for this study (e.g., Spatial Orientation), some tests developed by other groups, some tests designed by the aircrew selection research unit, and an early version of the Big 5 trait personality measure. Extensive analyses were conducted on these measures, including some comparisons between F-15 pilots and the basic trainees. However, only one set of analyses, lacking important descriptive statistics and correlational analyses, was published. Instead, partial correlations were reported between the individual predictors, taken one at a time, and situational awareness ratings, with flight experience statistically controlled for. Of the 6 predictor measures that showed a significant partial correlation, four were considered to be cognitive measures. These four tests (unit weighted) were summed and called the composite general cognitive ability. Only two psychomotor tests showed a significant partial correlation. These two tests were summed and called the composite psychomotor ability. A conscientiousness scale was also constructed from the personality measure. From a series of hierarchical regressions, the conclusion drawn was that only general cognitive ability, and not personality or psychomotor ability, added to experience in predicting situational awareness.

It is unfortunate that not more was done with this unique and rare dataset. The time of fighter pilots is a precious commodity, and getting 5 hours of it for them to present their cognitive, information-processing, psychomotor, and personality qualities is unlikely to be repeated. It would have been highly informative to have seen the results published for a factor analysis of the predictor measures, and a comparison of factor and item means

with various reference groups (non-fighter pilots; officer non-pilots; enlistees) as a way of discovering what "the right stuff" actually might be. If the data from this study could be located, there are numerous additional analyses that could be performed that could have a substantial impact on our understanding of the special qualities (perceptual, motor, temporal, and cognitive) of fighter pilots, and what differentiates the best fighter pilots from the very best. Systematic and comprehensive analyses of the data set could have had implications not only for personnel selection and classification, but also training, cockpit design, evaluating the effects of fatigue, drugs, and alcohol, and other psychological factors.

Carretta, T.R., Perry, D.C., Jr., & Ree, M.J. (1997). Prediction of situational awareness in F-15 pilots. *The International Journal of Aviation Psychology, 6(1)*, 21-41. **O-43**

Kyllonen, P.C. (2007). *The Learning Abilities Measurement Program (LAMP) 1982-1999*. San Antonio, TX: Operational Technologies Corporation. **E-03**

Project Combat Team
A Case Study of a Personnel Research Project
By C. Wayne Shore, Ph.D.

In the mid 1960s, Secretary of Defense McNamara sent a letter to the Secretary of the Air Force inquiring about the feasibility of replacing pilots with navigators in the second seat of the F-4. The Air Force response, based solely on subjective judgments of senior Air Force officers, was that a pilot was required in the second seat position. The DoD requested that the Air Force go beyond that initial response and conduct a study to develop data relevant to the issue. The Air Force generated a study and resulting data, but DoD analysts judged that the study lacked sufficient scientific rigor.

The Air Force was asked again to study the question, but this time with acceptable scientific protocols. The colonel in charge of the study assembled a team of four personnel research scientists, three of whom were currently assigned to the Air Force Personnel Research Laboratory and one who was a former member of that organization.

The team first met in February, 1968 at Davis-Monthan AFB, Arizona where they worked with F-4 aircrew instructors to define the tasks and responsibilities of the second-seat crewmember. The team designed the study, after which two team members went to Vietnam and two went to Thailand. There they interviewed F-4 crewmembers immediately following their sortie debriefing. Data were collected about what tasks the crewmembers performed and how well the tasks were performed. Some navigators had been assigned to second-seat duties, so that their performance could be compared to that of pilots in the second seat. Data collection was conducted in Southeast Asia over a period of several weeks.

The research team then assembled at Eglin AFB, Florida where they analyzed the data and prepared their report. They concluded from the data that there was no significant difference in performance between pilots and navigators in the F-4 second seat.

As a result of this study, navigators replaced pilots as second-seat crewmembers. The Task Force on Manpower Research estimated the "short-run" savings at $500M in avoided pilot training costs.

This study met all of the criteria for a strong personnel research program. Specifically the presence of an ongoing research agency provided the ready availability of a professionally qualified staff. Second, this study design applied technology that had been recently developed by the Air Force Human Resources Laboratory. Third, the return of investment, based on a savings of $500M, was overwhelmingly favorable. Fourth, this study demonstrated a rapid response capability. From the first meeting of the research team in February 1968 until its final report was submitted in September 1968, this major study was performed on a very timely basis. Finally, the study would not have been done without the issue being raised at high executive levels. This worthwhile study was the

result of a senior manager raising an important question and the Air Force having the capability of providing a scientifically sound answer.

The single most important lesson to be learned from this project is that Air Force managers should understand what issues can be constructively addressed by personnel research scientists and direct their activities accordingly.

Shore, C.W., Curran, C.R., Ratliff, F.R., & Chiorini, J.R. (1970, April). *Proficiency differences of pilot and navigator F-4 second-seat crew members: A Southeast Asia evaluation* (AFHRL-TR-70-9, AD-709 728). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **O-44**

Ratliff, F.R., Chiorini, J.R., Curran, C.R., & Shore, C.W. (1970, November). *Evaluating combat crew training performance using criteria of minimum performance standards* (AFHRL-TR-70-50). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **O-45**

Ratliff, F.R., Shore, C.W., Chiorini, J.R., & Curran, C.R. (1969, July). *Inflight performance differences of pilot and navigator F-4 second-seat crew-members: A limited Southeast Asia combat evaluation* (AFHRL-TR-69-104, AD-705 140). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **O-46**

New Content Areas for Officer Testing Program

**Methodology for Identifying Abilities
For Job Specialties (Project MIDAS)**

Military personnel testing officials at the Air Force Personnel Center have sponsored several efforts related to expanding the ability coverage of the AFOQT and rated selection test batteries. Among the follow-on proposals being considered are job analyses of officer specialties to provide a foundation for identifying ability areas for new test development. The task inventory approach to job analysis used at the Occupational Measurement Squadron provides comprehensive information about tasks performed by officers in different specialties. Further, psychologists with expertise in developing ability tests can use the task/job descriptive information from occupational surveys to support inferences about the underlying ability requirements for task performance. However, a more direct and empirically-based approach would be to build on methodological advances made in an Air Force research effort called Project MIDAS.

Project MIDAS, an acronym for Methodology for Identifying Abilities for Job Specialties, resulted in procedures for linking components of work in officer specialties with ability requirements. As one of only a handful of military efforts with this goal, the project is notable and worthy of further development and application. The methodology yields task-to-ability linkages allowing officer attribute requirements to be systematically defined. The process uses an Air Force ability taxonomy covering 28 domains (15 cognitive, 6 psychomotor, 7 interpersonal), task action verbs from occupational surveys (for example, "repair," "fly," "analyze," "interpret," and "inspect"), and expert judgments from Air Force supervisors about the importance of different abilities for task performance. A limited field test was completed with four officer specialties: Communications-Computer Systems Staff Officer, Information Management, Flight Safety Officer, and Pilot. An important finding was that supervisors agreed about the importance of different abilities for the task-verb work descriptions. With the reliability established, the methodology is ready for application in a broader sample of officer specialties to determine abilities which should be measured by officer selection tests.

Dittmar, M.J., Weissmuller, J.J., Driskill, W.E., Hand, D.K., & Earles, J.A. (1994). *Methodology for identifying abilities for job specialties (MIDAS)* (AL/HR-TP-1994-0008, AD-A277 919). Brooks AFB, TX: Human Resources Directorate, Armstrong Laboratory. **O-24**

Skinner, J. (2007). *New content areas for the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-25**

Leadership Effectiveness Assessment Profile (LEAP)

The Leadership Effectiveness Assessment Profile (LEAP) was a biodata instrument designed to measure specific traits predictive of Air Force officer leadership behavior. Development of the instrument proceeded using a conceptual model of officer effectiveness and retention, with major constructs derived from the literature, principally on leadership theory and empirical studies on officers. The experimental biographical survey instrument was designed to supplement the cognitive abilities measured by the Air Force Officer Qualifying Test (AFOQT) by tapping non-cognitive attributes judged to be important for officer performance and retention. Eight nonintellective constructs were chosen: Leadership, Commitment, Managership, Achievement Orientation, Adaptive Behavior, Socialized Power, Physical Fitness, and Retention Propensity. The instrument was rationally developed; items were written to correspond to the constructs. As the project progressed, several versions of the instrument were prepared and data collected on officer samples. Preliminary analyses addressed reliability and validity issues, as well as both rational and empirical scoring strategies for developing item keys for the constructs. Additional research on a faking detector scale for the LEAP was accomplished to address concerns about the susceptibility of the biodata survey to response distortion. A final version of the LEAP meeting psychometric quality standards was not achieved during the project, and the instrument was never used operationally.

Shermis, M.D., Falkenberg, B., Appel, V.A., & Cole, R.W. (1996). Construction of a faking detector scale for a biodata survey instrument. *Military Psychology, 8* (2), 83-94. **O-26**

Skinner, J. (2007). *New content areas for the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-25**

Officership

The military coined the term "officership," and it appears extensively in writings by and about the military. No widely accepted definition exists. The RAND Corporation explored using the term to describe a profession or occupational group for military officers that met the same standards for defining other professions like law or medicine.

```
                    ┌─────────────────────────────────┐
                    │  Characteristics of Officership  │
                    └─────────────────────────────────┘
```

Knowledge and skill expertise	Commitment and closed community	Competence	Formal code of law and ethics
• Military science • History • Management • Leadership • Specialization • Formal credentials	• "Calling" • Serve common good • Norms and values • "Total institution"	• Promotion based on potential • Centralized selection • Peer comparisons	• Guidelines for behavior and conduct • Sanctions

Figure. Defining officership as a profession

In contrast, an Air Force study of pilot selection used the term as a human ability construct --- an attribute possessed by an individual. Results showed that pilot selection boards members, using a "whole person concept," valued indicators perceived to reflect applicants' officership more highly than reliable measures of cognitive ability. The finding was perplexing in light of substantial research showing the predictive validity of cognitive ability for flying training outcomes. Analyses are limited, but to date officership measures used by selection boards have not been found to correlate with training criteria.

The need for research on officership has become a recurring theme in the past decade. Officership is seen as a potentially new construct for officer selection and an opportunity for experimental test development to complement mental ability measures in the AFOQT. If research is pursued, input from senior leaders will be critical for obtaining a consensus judgment about the meaning of the term and a basis for an operational definition. Components will need to be identified in order to develop reliable measures for an instrument like an officership assessment form. The extent to which personality traits, as measured by the Self-Description Inventory (SDI+), overlap with an "officership" construct is also of interest. Additional research will be required on appropriate validation criteria. Existing criteria of officer training performance are oriented to academic or occupational performance skills (for example, training grades or check flight scores) and are not likely to capture dimensions underlying behaviors or traits associated with officership.

Skinner, J. (2007). *New content areas for the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-25**

Weeks, J.L. (2000). *USAF pilot selection* (AFRL-HE-AZ-TP-2000-0004). Mesa, AZ: Warfighter Training Research Division, Air Force Research Laboratory. **O-27**

The Five-Factor Model of Personality

In the late 1950's, a landmark study by researchers at the Air Force Human Resources Laboratory, Dr. Ernest Tupes and Dr. Raymond Christal, found that five recurrent personality factors emerged from ratings on 35 personality traits taken from eight different samples. Personality theories had historically proposed a wide range of personality descriptors or traits with as few as two and as many as 20 separate personality dimensions. The Five-Factor Model succeeded in organizing personality descriptors under five unifying traits that appear to measure the basic dimensions of personality. Additionally, these traits were found across a variety of educational levels, ages, and cultures and under different administrative methods.

The Tupes and Christal study used peer ratings to assess 35 personality traits that were considered to be representative of the personality domain. The traits were first identified by Allport and Odbert and later by Cattell. The study consisted of peer ratings on these 35 personality traits taken on 8 samples. Three samples were from Air Force Officer Candidate School (OCS) and the ratings were from different-sized groups. One sample consisted of ratings by attendees at the Air Command and Staff School. Two were reanalyses of samples from Cattell's (1947, 1948) work, and two were reanalyses of samples from Fiske's work (1949). The analyses included peer ratings from people who were acquainted for as little as three days to as much as one year and who had as little as a high school education to graduate-level education. Some samples were Air Force personnel in various levels of training both enlisted and officer and some were university students. The type of rater ranged from naive persons to clinical psychologists and psychiatrists with years of experience.

The analyses of the different samples consistently revealed the same five dominant bipolar dimensions or factors.

1. **Surgency (also called Extraversion):** This factor is defined by the primary traits of Talkativeness, Frankness, Adventurousness, Assertiveness, Sociability, Energetic, Composed, Interest in Opposite Sex, and Cheerfulness.
2. **Agreeableness:** This factor is defined by the primary traits of Good-Natured, Not Jealous, Emotionally Mature, Mildness, Cooperativeness, Trustfulness, Adaptability, Kindliness, Attentiveness to People, and Self-Sufficiency.
3. **Dependability (also called Conscientiousness):** This factor is defined by the primary traits of Orderliness, Responsibility, Conscientiousness, Perseverance, and Conventionality.
4. **Emotional Stability (also called Neuroticism):** This factor is defined by the primary traits of Not Neurotic, Placid, Poised, Not Hypochondriacal, Calm, Emotionally Stable, and Self-Sufficient.
5. **Culture (also called Openness to Experience):** This factor is defined by the primary traits of Cultured, Esthetically Fastidious, Imaginative, Socially Polished, and Independent-Minded.

In 1993, Dr. Christal developed a computerized Self Description Inventory (SDI) to measure the Five-Factor Model using ratings on both traits (64) and behavioral statements (99). This rating inventory also resulted in a strong five-factor model and supported the findings of earlier studies. The United Kingdom then developed a paper-and-pencil version of the SDI called the Trait Self Description Inventory (T-SDI) which also proved reliable in predicting the five dimensions.

Findings from meta-analyses of the Five-Factor Model are revealing that each trait is valid, at least modestly, for prediction of some criteria and job groups. Conscientiousness has consistently been found to be valid for all criteria on all types of jobs. The meta-analyses show that personality dimensions can be valid predictors of performance, especially when the jobs are analyzed based on personality components and with a valid strategy for identifying predictors. Findings indicate that the measures have potential for increasing predictive effectiveness over the use of cognitive abilities alone.

The most recent research with Christal's SDI was completed under contract to develop experimental measures for future use in Officer selection. The SDI was modified to change single word descriptors to behavioral statements and lengthened to include scales relevant to Officer performance. The two new scales were Service Orientation to measure organizational skills and Team Orientation to measure propensity to work in groups rather than work alone. In 2005, this version, the SDI+, became an experimental addition to the operational Air Force Officer Qualifying Test (AFOQT).

Documentation of the Five-Factor Model was in government publications but received little visibility in the psychological literature until the 1980's when more psychologists recognized the five factors were fundamental to the measurement of personality. The significance of the Tupes and Christal work is that it clearly defined the five factors in numerous situations and showed them to be replicable and universal. There may be much to be gained by using non-cognitive variables such as personality traits to predict success in Air Force training and jobs. Factors other than job knowledge such as willingness to work and discipline which are also essential to job performance can be measured using the Five-Factor Model. Additional research is needed on social desirability responding, theories linking personality and performance, and matching personality attributes to jobs.

1992 *Journal of Personality*, Volume 60, Issue 2, pages 175-252 containing edited reprint of Tupes and Christal's (1961) technical report and with comments and introduction by R. R. McCrae (Editor). **O-28**

Barrick, M.R., & Mount, M.K. (2005). Yes, personality matters: Moving on to more important matters. *Human Performance*, *18*(4), 359-372. **O-30**

Tupes, E.C., & Christal, R.E. (1961) *Recurrent personality factors based on trait ratings.* (ASD-TR-61-97). Lackland AFB, TX: Personnel Laboratory, Aeronautical Systems Division. **O-29**

Self-Description Inventory (SDI+)

The Air Force has a history of personality test development extending back to the mid-1950s. Results of this early work led to the identification of five recurring personality factors: Conscientiousness, Agreeableness, Neuroticism, Openness, and Extraversion. Later research outside the military verified the ubiquitous nature of these factors across a broad range of personality tests and subject populations, and the factors became known as the "Big Five."

In the 1990s the AFHRL sponsored research to construct a contemporary inventory suitable for computer administration. The inventory was called Christal's Self Description Inventory, and it used both single-word trait adjectives and more lengthy behavioral statements to measure the "Big Five." Beginning in 2000, additional research was completed under contract on experimental measures for the officer testing program. An objective was to bring a "Big Five" personality test nearer to operational implementation and to extend the traditional measures in new directions by measuring additional traits deemed relevant to officer selection. Christal's personality inventory was modified to change single-word descriptors to behavioral statements and lengthened to prepare the Self-Descriptive Inventory (SDI+).

The SDI+ personality test has seven scales, five for measuring the "Big Five" personality traits and two scales (Service Orientation and Team Orientation) for assessing desirable characteristics of military officers. The Service Orientation and Team Orientation scales align with senior leaders' perceptions about performance requirements for officers. The Service Orientation test was designed to capture an officer applicant's potential for organizational commitment prior to service entry. Test development in this area supported the broad Air Force goal of fostering "professionalism" versus "careerism" among officers. The Team Orientation test was designed to assess predispositions for working comfortably in groups versus preferences for working alone.

An initial try-out of the SDI+ was completed with a sample of basic airmen. Results of a factor analyses with a 7-factor solution indicated that six of the seven scales were independent. The Team Orientation scale was not completely independent of the "Big Five" scales. However, it was retained in the inventory pending results from additional field testing and validation with officer samples.

In 2005, the SDI+ became an experimental adjunct to the operational AFOQT. The inventory has 220 items and requires 40 minutes to administer. Scale scores are not used for selection decisions but data on officer applicants are being compiled to support additional research. Research issues include measurement stability and concurrent and predictive validity for officer performance measures. Response coachability and faking will be major issues if the SDI+ is used for selection but will be of less concern if it is applied only for improving job-fit in officer job assignments.

Alley, W.E. (2002). *Development of experimental measures for the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-06**

Substituting Commercial Tests for the AFOQT

The feasibility of using commercial tests instead of the AFOQT, which is developed, administered and maintained by the Air Force for officer selection, has been addressed on numerous occasions. Discussions focused on tradeoffs between the AFOQT and college admissions tests (SAT Reasoning Test and ACT) for ROTC cadet selection and commercial graduate school admission tests (like the Graduate Record Examination) for college graduates applying for OTS. HQ AFROTC has been the principal proponent for eliminating the AFOQT and substituting commercial tests.

Relevant research addresses the similarity in ability measurement of the AFOQT and commercial tests and the validity of national standardized tests for predicting military performance criteria. Several studies show that test-takers' scores on the AFOQT and SAT/ACT are correlated. Further analyses of verbal and quantitative composite scores of the AFOQT and SAT reveal that the tests assess similar abilities and are construct equivalent for measuring general mental ability. However, the tests differ in difficulty. Lack of parallelism in score distributions indicates that the tests are not directly interchangeable for the purpose of making personnel selection decisions and that validity results for the AFOQT cannot be assumed to generalize to the SAT. Validation studies of the SAT itself are limited in number and are not available for current samples of officer applicants. Older studies, however, suggest the correlational patterns for selected performance criteria for ROTC cadets are similar for the AFOQT and SAT. Comparable analyses have not been completed for the ACT or the GRE against relevant measures of officer performance.

Answers to questions about using the SAT and GRE in lieu of the AFOQT are not clear-cut. Studies by the Educational Testing Service show that both commercial tests are reliable cognitive ability measures. In general, based on their psychometric properties, there is nothing to preclude their use. Further, Air Force data suggest that due to similar verbal and quantitative measurement properties, the SAT could be substituted and most likely would not practicably impact predictive validity for cadet training program outcomes. However, the advantages of making the substitution are not apparent. In the past, concerns raised by HQ AFROTC did not appear to be with the AFOQT *per se* but with selection standards set on the test metric. Whether cognitive ability is measured with the AFOQT or a commercial test, entry standards will continue to be imposed on applicants for commissioning to insure a capable officer force. Further, a decision by the Air Force to lower standards to address HQ AFROTC concerns about detachment viability and pilot training qualification rates can be accomplished with either the AFOQT or a commercial test.

In addition to entry standards, managers would need to consider numerous policy implications and tradeoffs. With commercial tests, the Air Force would lose control over decisions about test content, item difficulty, available of scores, testing schedule, and retest policies. They would still have the obligation of meeting national standards by conducting predictive validity, test bias, adverse impact and standard setting studies to

defend use of the tests for military officer selection. Other issues are potential savings in AFOQT test development costs and testing time. The advisability of using different commercial tests (SAT and GRE) for AFROTC and OTS commissioning programs would have to be addressed, along with considerations about tracking and comparing officer quality from different commissioning sources.

Applicants applying for rated training pose another set of issues. One of the difficulties of a one-to-one swap with a commercial test is that the AFOQT verbal and quantitative subtests are also used in the Pilot and Navigator-Technical composites. Presently, there are no analyses showing whether the SAT or GRE verbal and quantitative scores could be incorporated into the Pilot and Navigator-Technical composites without appreciable loss in predictive validity for rated training criteria. The advisability of using applicants' SAT scores from their junior and senior years in high scores for selection decisions for pilot training five to six years later needs is questionable. Further, five other AFOQT subtests are presently scored in one or both of the rated selection composites (Instrument Comprehension, Block Counting, Table Reading, Aviation Information, and General Science). The Air Force would need to address test development activities and costs for these or similar content areas using either paper-and-pencil or computer administration, for example, as part of the platform currently supporting the Test of Basic Aviation Skills (TBAS) for pilots.

Ingerick, M. (2005). *Identifying leader talent: Alternative predictors for U.S. Air Force junior officer selection and assessment* (DFR-05-47). Alexandria, VA: Human Resources Research Organization. **O-31**

Skinner, J. (2007). *Substituting commercial tests for the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-32**

Officer Commissioning Programs

AFROTC Detachment Effectiveness Measurement

In 1965 the Department of Defense requested that the services develop procedures for evaluating ROTC detachment effectiveness and determine whether certain detachments should be disestablished. The request and the series of Air Force studies which followed were conducted in the turbulent years of the late 1960s and early 1970s. The gradual phasing out of conscription, the progressive elimination of compulsory ROTC programs, and a sharp drop in officer requirements were combined with a hostile environment on college campuses characterized by anti-Vietnam War and anti-ROTC protests. The prospects for ROTC seemed grave, and many Congressmen and defense department officials began to question ROTC's viability as the premier commissioning sources for the armed forces. Originally the sole criterion for disestablishment was number of graduates per year from each detachment. Air Force researchers pointed out that a more realistic criterion would be cost per graduate and that detachments differed in other ways, many of which should be considered before abolishing a program.

The Air Force Human Resources Laboratory was asked to define criteria for assessing the effectiveness of Air Force detachments and to develop longitudinal data bases for use by ROTC program managers. Measures included: 1. *production criteria* (number of graduates, number entering active duty); 2. *rated training criteria* (number of entrants, graduates, eliminees); 3. *retention criteria* (number of active/inactive graduates for rated and non-rated specialties); 4. *aptitude, quality and officer effectiveness criteria*; 5. *cost effectiveness criteria* (cost per graduate by type, e.g., pilots, navigators); and 6. *college characteristic variables* (measures of host colleges including selectivity, professional orientation, student body size, geographical location).

As the research program progressed, several improvements in measurement procedures were attempted. Grade point averages were adjusted for college selectivity to obtain more comparable measures across detachments. Average officer effectiveness reports were adjusted for yearly inflation and rating form differences. The data base was shown to be a reliable and accurate management tool for evaluating current detachments, predicting the viability of proposed sites based on host college characteristics, and simulating policy changes.

Findings from the completed studies are dated and of little value for assessing present detachments. However, the measurement strategies and analytic techniques produced robust results. Procedural details may be of interest to managers as they respond to an apparent post-9/11 attitude shift and growing support among college administrators and students for a return to university-sponsored officer training, even at Ivy League universities which abolished ROTC programs during the 1960s.

Skinner, J. (2006). *Air Force Reserve Officer Training Corps (AFROTC) studies*. San Antonio, TX: Operational Technologies Corporation. **O-34**

Weighted Factors Selection System for ROTC Professional Officer Course

The Professional Officer Course (POC) is designed to prepare cadets during their junior and senior years in college for officer commissioning. It is administered at AFROTC detachments located on college campuses. When the quality of cadets decreased after the all-volunteer force policy, a weighted POC selection system (WPSS) was implemented in 1977. Cadets selected in the next two years had higher standardized test scores and grade point averages (GPA).   A follow-up validation study showed applicants with higher WPSS scores performed better in POC, in later technical training courses, and on the job.

The WPSS was developed by HQ AFROTC and Air University with analysis support from the Air Force Human Resources Laboratory.   The weighted selected system captured the consensus policy of board members concerning what factors were important for selecting candidates for the POC and the relative contribution of each factor.   The 1977 policy capturing process resulted in eleven factors which were differentially weighted and then combined to produce an overall measure of applicant quality called the Quality Index Score (QIS). The table shows the factors for the original QIS.

Factors in POC Selection Systems (1977 – 2006)

Factors	Original 1977 QIS	1978- 1982 QIS	1988 QIS	1996 QIS	2006 OM
AFOQT Academic Aptitude	X	X	X	X	
Scholastic Aptitude Test (SAT total)	X	X	X		X
Cumulative GPA	X	X	X	X	X
Detachment Commander Rating	X	X	X	X	
College Selectivity Rating	X				
AFROTC GPA	X	X			
AFOQT Quantitative	X	X	X		
Type Program (credit for mil. courses)	X	X			
Academic Major (technical major credit)	X	X			
Total Number of Applicants/Cadets	X	X			
Applicant/Cadet Rank	X	X			
AFOQT Verbal			X		
Physical Fitness Test				X	X
Relative Standing Score (combines Unit Commander Ranking & POC class size)					X

The table shows subsequent changes made to the QIS by HQ AFROTC and the factors used to compute an Order of Merit (OM) in the current selection system.   The early WPSS factors differ markedly from the selection formula used today. Major changes were decreased use of the AFOQT for cognitive ability measures and increased emphasis on cadets' physical fitness. A validation study is needed for the current system.

Skinner, J. (2006). *Air Force Reserve Officer Training Corps (AFROTC) studies*. San Antonio, TX: Operational Technologies Corporation. **O-34**

AFOQT-SAT-ACT Conversion Tables

The HQ AFROTC has used conversion tables for the AFOQT, SAT, and ACT tests for at least the last three decades in their procedures for selecting cadets for the Professional Officer Course (POC). The tables allow cadet's observed scores on one test to be used to estimate their performance on the other test(s). These score conversion tables have often been a source of controversy.

The Air Force Human Resources Laboratory (AFHRL) was asked on several occasions to conduct the test equating research; sometimes the lab cooperated and sometimes the lab refused. The inconsistent policy arose because AFHRL wanted to be responsive to customer requests on one hand, but knew the tests were not parallel and had concerns about the equating accuracy on the other hand. Another issue was abundant evidence that the tables were used improperly in the field. For example, AFHRL conversions were usually one-way, allowing AFOQT scores to be used to estimate SAT or ACT scores. However, in the past, it was clear that the tables were used operationally for reverse conversions, despite repeated warnings about the error introduced from improper use. On other occasions, Air Force organizations (for example, Recruiting Service) contracted directly with the ACT or Educational Testing Service to develop equating tables for the AFOQT and the college admissions tests.

The central issue underlying the conversion tables is that the tests are not parallel. The AFOQT is a more difficult test, and the shapes of score distribution for the tests are sufficiently dissimilar to call into question the defensibility of the score conversions. The most recent research on the topic clearly shows the problems that can be encountered. Using an equipercentile equating method, scores on the SAT were rescaled to the AFOQT scale, and impact analyses were conducted. Significant differences were found in qualification rates for male vs. female and African-American vs. White subgroup comparisons using equated-SAT to AFOQT scores. The single equating worked to the advantage of some groups but not to others, an unacceptable outcome for use of a test in personnel selection decisions. The equating was not sufficiently general, robust, and accurate to be used for all groups. The study showed that separate equatings, one for each sex/race subgroup, would be necessary to yield equal qualification rates. The use of differential scoring (equatings) for race/sex groups is prohibited in selection systems by the Civil Rights Act of 1991.

The research has implications for the tables currently used by HQ AFROTC for POC cadet selection. Cadets' observed test scores should be used instead of equated scores whenever possible.

Ree, M.J., Carretta, T.R., & Earles, J.A. (2003). Salvaging construct equivalence through equating. *Personality and Individual Differences*, 35 (6), 1293-1305. **O-35**

Skinner, J. (2007). *Substituting commercial tests for the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-32**

Grade Point Average as an Officer Selection Factor

In the field of personnel selection, the situation often arises where applicants from different universities are in competition for the same position. If undergraduate grade point average (GPA) is a selection factor, an important issue is whether to take into account the effect of possible college differences in the meaning of GPAs. How does a GPA from a large state supported school compare to a slightly lower GPA obtained from a small prestigious private college? Should a selection official make allowances for the presumed quality of the school attended or does a 3.5 GPA from any college equate to the same level of expected training or job performance? These questions are relevant in the Air Force, because GPA is considered by selection boards for officer training programs.

There is a voluminous body of research demonstrating that GPA is a valid but modest predictor of later training and job performance. Findings from Air Force studies are consistent with those from the private sector. GPA has a significant relationship to performance criteria but the validities are small, especially in comparison to those for cognitive ability tests like the AFOQT. In more refined analyses of private sector studies, validities for graduates of the same college were found to be higher on average than those for graduates of different colleges. These findings were suggestive, but they did not directly address the question of whether GPAs from different colleges have the same meaning in terms of future expected performance levels for job applicants or officer applicants.

Researchers at AFHRL were in a unique position to tackle the question, because of their access to archived data on large samples of officer candidates. They found a joint college and GPA effect consistently for measures of cadets' academic performance in OTS. Cadets who had the same GPA but who had graduated from different colleges performed differently. Follow-on analyses of explanatory factors used measures of the characteristics of colleges attended. About 20 to 40 percent of the variance in expected performance levels due to colleges was explained by college selectivity (average college admission test scores for freshman classes and the selection ratio (percent of applicants selected)).

Empirical support was obtained for the widely held belief among personnel selection officials that grades vary in meaning across colleges, and that a "C" at one college may be equivalent to an "A" at another college. The GPA is not a common yardstick. Agencies like the Air Force which use standardized test scores from the AFOQT in addition to GPA may capture some of the performance variance due to college differences.

Romaglia, D.L., & Skinner, J. (1990). Validity of grade point average: Does the college make a difference. *Proceedings of the 32nd Annual Conference of the Military Testing Association*, p. 345-350. Orange Beach, AL. **O-33**

Officer Performance Appraisal

Studies of the Officer Effectiveness Report (OER)

Procedures for evaluating officers' performance and promotion potential have changed numerous times, and characteristically the process has yielded inflated ratings. Early research addressed two broad areas: 1) the OER as a personnel management tool, and 2) the OER as a research criterion measure. The acronym OER is used, because the studies summarized predate the adoption of the present Officer Performance Report (OPR) terminology. Whether the research findings from the first two areas generalize to the present OPR system is unknown due to changes in rating scales and procedures. The nature of the studies reveals the breadth of interest in the OER. Some research questions and methodologies are still relevant but studies with updated data bases of OER ratings would be needed to insure currency of results.

In a 1968 review of a decade of research on the OER as a personnel management tool, it was noted that numerous studies involving descriptive and inferential analyses were completed for situational and demographic variables. The investigations focused on questions about the nature and extent of relationships between OER ratings and officers' grade levels, command and AFSC, where significant differences were often found. The results were mixed about whether the differences were attributable to systematic selection of more capable officers for higher grades and responsibility. A confounding factor was the observation from trend analyses of major shifts in mean ratings with the introduction of new rating forms. Mean ratings would decrease shortly after new procedures were implemented but indications of rating inflation would soon reemerge. Studies of educational effects did now show a consistent increase between educational level and performance ratings.

The second broad area of research explored the utility of the OER as a criterion measure of officer performance. These studies were concerned with the measurement and improvement of officer selection devices and training programs. A consistent finding was that cognitive measures were not useful predictors of officer effectiveness ratings. The lack of relationship was attributed to restriction in range from prior selection of officers on ability tests and rating inflation which reduced variance in the OER measure. Other studies on the predictability of physical proficiency tests, athletic ability, and biographical data reported, with few exceptions, near zero relationships with the OER criterion. Non-cognitive measures such as personality trait ratings and peer ratings showed greater promise for predicting officer effectiveness reports than cognitive measures. Early studies also reported that the OER was the most important variable in accurately estimating the promotion board score.

Improvements in the accuracy of the officer rating procedures would increase the utility of the OER as a personnel management and assessment tool.

Sturiale, G. (1968). *The officer effectiveness report as a performance measure: A research review* (AFHRL-TR-68-113). Lackland AFB, TX: Air Force Human Resources Laboratory.
O-36

Controlled Promotion Potential Distribution

In 1968, efforts were initiated to develop a new officer evaluation system but no agreement was reached among the Air Force Council and major air commands on how new forms should be designed. In 1971 the AFHRL was asked by HQ USAF, Director of Personnel to develop a research-based solution for a new OER. The development effort was overseen by a panel of general officers appointed to serve as an OER Review Group. Consultants were brought in from academia, industry, and other military Services.

The major features proposed by AFHRL included separate forms for evaluating performance and promotion potential. The performance evaluation was structured around a statement of job objectives to be determined by the ratee and rater. Nine performance factors were identified based on several studies involving analyses of OER word pictures, frequency of use, and importance. After testing several rating scales, a five-point behaviorally-anchored scale based on standards of performance (ranging from Far Below Standards to Well Above Standards with Meets Standards at the midpoint) was chosen to rate officers on the performance factors. The promotion potential evaluation consisted of a 3-point scale summarizing the rater's assessment of the officer's overall performance evaluation (Does Not Meet Standards, Meets Standards, and Exceed Standards). In addition, the **rater** was to provide a promotion recommendation of the ratee's potential for positions of greater responsibility, compared with officers of the same grade, on a 7-point scale (Retain in Present Grade, Lower 1/3 in Primary Zone, Middle 1/3 In Primary Zone, Top 1/3 in Primary Zone, 1 Year Ahead of Year Group, 2 Years Ahead, 3 Years Ahead). This proposed system, which was thoroughly grounded in research, was transferred to HQ USAF.

HQ USAF called for a review by military officers, made significant modifications to the proposed system, and implemented a revised OER system in November 1974. Some of the features proposed by AFHRL were retained; others were not. One major modification was to the Evaluation of Promotion Potential section of the OER. The scale recommended by AFHRL for the rater to complete was dropped. In its place, six blocks were used, with the bottom block labeled lowest "potential" and the top block labeled highest "potential." Scaling referents for the intermediate blocks were not defined. The rater, additional rater, and reviewing official each entered a promotion potential rating by recording an "X" in one of the six blocks. Particularly noteworthy was the use of a **controlled** distribution which constrained the **reviewer** only to assigning a maximum of 22% of ratings to the top block, a maximum of 28% to the second block, and allowed the remaining 50% to be distributed among the remaining four blocks. The **controlled OER** was a quota system. The controlled system was being used effectively in other private and public sector organizations to force differentiation among ratings. However, Air Force officers strongly resisted the controlled OER, and it did not last long. By 1977, the 28% limitation on second block ratings was removed, and by 1978 the controlled OER era had ended. The controlled OER was widely perceived to be a mistake in a culture where top marks for officers suddenly became the exception rather than the rule.

In 1989, a review of the officer evaluation system by Syllogistics Inc. and The Hay Group resulted in several conceptual designs to provide differentiation in potential for promotion ratings. One was similar to the former controlled OER. However, recognizing officers' negative attitude toward the prior attempt to control rating distributions, the proposal was a modest ten percent target for early promotion ratings.

In lieu of forced distributions or other methods for distinguishing the capabilities of officers for higher grades and responsibility, the discriminating factors for promotion decisions in the Air Force are the rank of the indorsing official and his/her narrative remarks.

Bottenberg, R.A. (1978). *Relationships among factors in new officer effectiveness report system* (AFHRL-TR-78-40). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-37**

Johnson, C.A., Meehan, J., & Wilkinson, R.E. (1976). *Officer effectiveness report development – 1971 through 1972* (AFHRL-TR-76-61). Lackland AFB, TX: Air Force Human Resources Laboratory. **O-38**

Nataupsky, M., Curton, E.D., Waller, E., & Ratliff, F.R. (1977). *Report on the 1975 officers' OER opinion survey* (AFHRL-TR-77-37). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-39**

Syllogistics Inc. and The Hay Group (1987*). Final report: Air Force officer evaluation system project* (DTIC Report No. AD-A223 834). Springfield, VA. **O-40**

Officer Grade Requirements

The Air Force procedure for determining officer grade structure for many years has responded to career planning factors and to the awarding and controlling of pay. The determination of grade is not clearly based on the requirements of the job. In 1963, the Vice Chief of Staff of the Air Force announced a study to develop a scientific system for determining officer grades that was to be conducted by the Air Force Human Resources Laboratory (AFHRL).

A comprehensive approach to determining grade requirements combined proven job analysis techniques with policy-capturing procedures, both of which were developed at ARHRL. It became one of the largest job analysis studies ever conducted in the military or civilian arena. The study was divided into three phases: 1) obtaining policy decisions based on ratings of members of a Policy Board about the appropriate grades for a selected sample of jobs, 2) developing policy equations to predict grade ratings given by the Policy Board to jobs in the sample, and 3) application of the policy equations to jobs remaining in the Air Force population.

Job descriptions were collected from 79,750 officers. From this sample, 3,575 job descriptions were selected for the criterion sample. A Lieutenant General led a panel of 22 Colonels who represented all commands. The Colonels read the descriptions and recommended a grade level for each position without knowing what the actual grade level was for the position. Five panel members rated each description There was high agreement among the board members on the grade ratings and board members expressed confidence in their ratings. Board members did not show bias toward jobs in particular commands or job specialties. They did, however, agree that many Air Force jobs were inappropriately graded and analysis showed that there was strong agreement among the raters on how the jobs should be upgraded or downgraded

Using the ratings given by the policy board, a mathematical equation known as a policy equation was developed that could predict the ratings of the board. More than a hundred predictor variables were first entered into the equation, but nine variables proved to be the most efficient at predicting the policy board ratings. Some of the variables were available from the job description information but some of them had to be rated by people in the field. The variables found to be most predictive of the policy board ratings were Management, Planning, Special Training and Work Experience, Judgment and Decision Making, Communication Skills, Level of Organization in which the job occurs, Mean Grade Rating by Field Judges, and Supervisor's Judgment of Appropriate Grade. The correlation of these variables with the policy board ratings was .84. After the policy equation was developed and demonstrated to predict the policy board ratings, it was applied to 10,000 additional job descriptions that were not rated by the policy board and found to be highly stable.

Finally, the data were used to make projections to the Air Force population except for lieutenants and captains since their grades were not controlled, general officers, doctors and dentists, and assorted other small groups. In every utilization field, the findings

showed some positions needed to be upgraded and some needed to be downgraded. A significant number of jobs needed to be downgraded in the pilot and navigator-observer fields. Making these adjustments would have required a significant change in aircrew management practices. The study also showed that the Air Force was somewhat undergraded at the Colonel and Lieutenant Colonel levels and considerably undergraded at the Major level.

Benchmark scales were developed to measure each individual officer description. Ten job factors were rated for each officer description. Each job factor was rated on a scale of 1 to 9 and each point in the scale was anchored with 3 descriptions of jobs that would be performed at that level. The benchmark procedure yielded a validity of .90 with Policy Board Decisions.

In 1974, the Air Force asked for a technology by which Management Engineering Teams (METs) could determine the appropriate grade requirements of all officer positions except line pilots, navigators, physicians, dentists, and personnel not subject to the constraints of the Officer Grade Limitations Act. The approach was to have the METs rate the jobs using the benchmark scale and tie the ratings back to the original Policy Board Ratings. The results showed that METs could accurately implement the policy of the 1964 Board.

The OGR studies that began in 1964 with additional analyses in the 1970's and 1980's resulted in ratings and recommended grades for 23,000 jobs and projections addressing 176,000 officer jobs. Initial studies included aircrew and non-aircrew positions, but issues with the management of aircrew positions led to later studies that included only non-aircrew positions. The Officer Grade Requirement study used a scientific approach to effectively establish a valid procedure for identifying the grade requirements for officer positions based on the content and responsibility of Air Force jobs. The methodology also had a dual purpose in that could be used to evaluate individual jobs and job grades could be compared across specialties.

Although the OGR methodology was not implemented by the Air Force, the approach was used to respond to a GAO query about why the proportion of officer positions to enlisted positions and average level of the positions exceeded those of the other Services. The second Air Force study justified the requirements and results were provided to the GAO. The Air Force must use grade authorizations to meet career planning objectives, but these needs could be evaluated along with recommendations based upon job requirements to achieve the best structure for officer grades.

Christal, R.E. (1975). *Systematic methods for establishing officer grade requirements based upon job demands* (AFHRL-TR-75-36, AD-A015 756). Lackland AFB, TX: Air Force Human Resources Laboratory. **O-41**

Finstuen, K., Matthews, G.N. & Pope, W.H. (1980). *Management engineering team applications of officer grade requirements method* (AFHRL-TR-80-32, AD-A093 508). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-42**

III. METHODOLOGIES FOR ADDRESSING

AIR FORCE PERSONNEL PROBLEMS

Hierarchical Grouping

Hierarchical grouping is a technique for grouping a set of regression equations to minimize the overall loss of predictive efficiency at each stage of clustering. Loss in predictive efficiency is measured by the decrease in overall squared multiple correlation coefficients. The technique was developed by the Air Force Human Resources Laboratory beginning in the 1960s, and was widely applied in Air Force research programs. There were no solutions available in the mathematical and statistical fields for addressing the unique grouping and clustering needs of the military. The hierarchical grouping method was developed to address shortcomings in statistical approaches such as factor analysis, discriminant function analysis, pattern analysis, and cluster analysis.

The approach is based on the concept that items should be grouped in an iterative fashion so as to maximize payoff or minimize cost at each stage in terms of some relevant criteria. An important feature is that the criterion function to be optimized is selected by the investigator or policy maker and can be varied from research problem to research problem. Using multiple linear regression techniques, a separate equation is computed for each criterion to be maximized or minimized. For example, the criterion could be grades for a single Air Force technical school or one policy maker's ranking of airmen on promotability. In these cases, the predictors could be scores on several aptitude tests related to training performance or scores on military experience and promotion test factors related to promotability. At this initial stage in the process, there are as many regression equations as there are technical schools, possibly hundreds, or as there members on a promotion board, possibly a dozen or two. To accomplish the needed reduction of the separate equations, hierarchical grouping combines the most similar regression equations iteratively. At each successive stage, the number of criteria (equations) is reduced by one. And, at each stage, the assignment of criteria (equations) into a given number of clusters makes the most accurate overall prediction of scores in this number of clusters. The quantitative measure of predictive efficiency lost at each iteration is the squared multiple correlation coefficient. Moreover, the hierarchical grouping method provides an optimally-weighted predictor composite for estimating scores on each separate criterion and cluster in the array at each iterative stage. In the final grouping stages, where the number of clusters becomes increasingly smaller, the analyses will begin to reveal clusters or groups of job families in which training performance for several technical schools is predictable from similar patterns of aptitude tests or similar policies for board members about factors important for promotion decisions. The equation(s) for the final cluster(s) provides information, through the size of regression weights for each predictor, about identifying aptitude tests for job family composites or weighting the relative importance of factors judged to be important for a promotion system.

The first application was in the development of personnel assignment programs to group families of Air Force specialties requiring similar aptitudes. Hierarchical grouping allowed regression equations relating aptitude predictors and technical school success criteria to be grouped to define job families that minimized loss of differential classification effectiveness in aptitude test batteries. Later the technique was used to

empirically verify the configuration of the Mechanical, Administrative, General, or Electronic (MAGE) aptitude composite structure. Other major applications were in developing enlisted promotion systems. Hierarchical grouping was used in conjunction with the policy capturing technique to design the Weighted Airman Promotion System (WAPS) and the Senior NCO Promotion Program (SNCOPP). The hierarchical grouping method was also used in studies of training priority based on task emphasis ratings, the structure of maintenance jobs, and time to cross-train among specialties.

In 2006, the hierarchical grouping software written by AFHRL for a mainframe computer was updated to run on a personal computer. The software upgrade was essential for an ongoing study of Air Force aptitude composites. Composite validity and classification efficiency are being examined in response to changes in the subtest content of the Armed Services Vocational Aptitude Battery (ASVAB). The Numerical Operations and Coding Speed subtests have been dropped and an Assembling Objects (AO) subtest has been added. Newly configured composites for the Air Force enlisted classification system are needed. Hierarchical grouping will be part of the analysis plan to account for structural changes in the enlisted classification system, to incorporate effects of technology changes in the initial specialty course content, and to address aptitude changes required to predict training performance.

Bottenberg, R.A., & Christal, R.E. (1961). *An iterative technique for clustering criteria which retains optimum predictive efficiency* (WADD-TN-61-30). Lackland AFB, TX: Personnel Laboratory. **RM-01**

Treat, B.R. (2007). *Hierarchical grouping in Air Force personnel analysis.* San Antonio, TX: Operational Technologies Corporation. **RM-02**

Ward, J.H., Jr. (1961). *Hierarchical grouping to maximize payoff* (WADD-TN-61-29, AD-261 750). Lackland AFB, TX: Personnel Laboratory. **RM-03**

Policy Capturing

Policy capturing is a decision-making model developed by AFHRL researchers in the early 1960s. Two of the best known applications, the Weighted Airman Promotion System (WAPS) and the Senior NCO Promotion Program (SNCOPP), were implemented.

The model is designed to "capture" and quantify the policy of a single rater or judge or of multiple raters or judges. The multiple-rater situation often arises in the Air Force where a policy board process is used for personnel selection, assignment, and promotion decisions. The mathematical technique associated with the policy-capturing model is multiple regression analysis, which is used to identify the variables (factors) considered by the board and to determine how these variables must be weighted to reproduce the board's actions. If there is high agreement among the raters and judges, a consensus policy can be "captured" from the regression equations using a hierarchical grouping technique to arrive at a single policy. Further, if more than one policy exists among board members, each policy can be identified and described with the model, and differences in policies brought to the attention of the raters or judges for arbitration.

To illustrate the policy-capturing model, its application for developing the WAPS will be briefly described. The objective of the project was a mathematical model that expressed or "captured" the consensus judgment or "policy" of highly qualified and experienced military personnel about the relative merits of airmen eligible for promotion. Since promotions had previously been based on the recommendations of promotion boards, the policy-capturing technique identified the optimum variables to be considered in the promotion formula based on the policies that the board members used in ranking airmen for promotion. First, an experimental promotion board composed of several members or judges was convened to rank-order a random sample of eligible airmen according to their promotability. Each airman's record displayed numerical values for his/her performance on the promotion variables. Each experimental board member was required to review records for all airmen and to make an independent judgment as to their rank order from most promotable to least promotable. Second, using multiple linear regression techniques, a separate equation was computed for each member of the board. The airmen promotion factors served as predictor variables and the ranks assigned as the criterion. The separate equation for each board member represented his/her promotion policy. At this stage in the process, there were as many regression equations as there were board members. Third, the multiple equations were reduced to a single consensus equation. To accomplish the reduction, a criterion-grouping technique, referred to as hierarchical grouping, was used to combine the most similar regression equations or promotion policies in an iterative process until there was only one common policy representative of all or the majority of the raters or board members. The final equation provided information, through the size of regression weights for each factor, about the board's judgments concerning the relative importance of each factor to promotion. The policy captured and implemented was the factors and weights used in the present WAPS.

Most applications of the policy-capturing model by AFHRL researchers were to personnel-management problems where "people" were the focus of analyses. These studies included modeling officer promotion boards, officer selection boards, and pilot selection boards; developing and implementing a weighted-factors selection system for the HQ AFROTC Professional Officer Course; determining the relative importance of certain variables in accounting for the proficiency of airmen working in particular career ladders; developing promotion policies for civilians; and simulating assignments of pilot trainees into specialized aircraft training tracks (fighter, bomber, tanker, transport). Other studies were conducted where the focus of the modeling effort was not on people. The largest of these efforts was the Officer Grade Requirements (OGR) project, in which the model was used to determine the appropriate distribution of grades for jobs in various officer specialties and utilization fields. Procurement managers' decisions about supporting or not supporting research and development projects have also been simulated.

Policy capturing is a widely applicable and quantitative way of taking a fuzzy decision-making process and making it well defined and replicable. In the context of promotion systems, it has the advantage of making the factors relevant for promotion visible to the personnel most affected by the process. Military managers and researchers should be familiar with the technique.

Christal, R. E. (1967). *Selecting a harem and other applications of the policy-capturing model* (PRL-TR-67-1). Lackland AFB, TX: Personnel Research Laboratory. **RM-04**

Fast, J.C., & Looper, L.T. (1988). *Multiattribute decision modeling techniques: A comparative analysis* (AFHRL-TR-88-3). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **RM-05**

Policy Specifying

The judgment process called policy specifying was developed by Dr. Joe H. Ward, Jr., a mathematician at AFHRL, in response to a requirement for a computer-based job assignment system for enlisted personnel. The system required a procedure to generate a "payoff" or "value" of the assignment of each recruit for each possible job (Air Force specialty). The problem was that the "payoff" values were unknown. Policy specifying provided a process for translating into mathematical form a policy maker's natural language statements about the general properties that a model of "payoff" should have. The policy specifying model was used to develop both pre-enlistment and post-enlistment person-job matching systems for the Air Force.

The decision maker, with the assistance of an analyst, decides upon the decision objective, for example, the value or payoff of a particular person-job match. The variables or attributes important in making the decision are defined (filling of quotas, maintaining minority balance, aptitude and interests of the recruit, learning difficulty of the job). Decisions are made about which pairs of variables should logically interact, and functional relationships between the variables are defined. This process continues in iterative and hierarchical fashion moving up through the levels of interacting variables until the overall decision objective is mathematically specified.

An integral part of the process is that as various functional forms between variables are modeled, the output is shown to the decision makers. Modifications are made until the decision maker is satisfied that the output of "payoff" values is consistent with their policy. An advantage of the policy specifying process is that it guides managers and policy makers in expressing potentially complex functional relationships between variables that may include interactions and non-linear forms.

This benefit is illustrated on the next page. While there are several variables that contribute to the expression of the value or worth of assigning a particular airman to a particular job, an important component involves two basic variables – aptitude of the person and difficulty of the job. Using the policy specifying approach, Air Force managers decided that on a "payoff" scale from 0 to 100, the value of placing airmen with differing aptitudes (40th to 95th percentile scores on the ASVAB) into jobs of different difficulty (entry aptitude requirement from 40 to 100) varied in the manner shown in the figure. Payoff is near zero for placing low aptitude personnel in high difficulty jobs. Payoff increases for airmen with higher aptitudes and as jobs become more difficult. As job difficulty increases, the amount of change in the value or payoff per unit change in airman aptitude increases rapidly. Payoff increases as a function of per unit change in airman aptitude depends on the level of job difficulty. The largest increases in payoff occur when both aptitude and difficulty are high.

The policy specified in the manner describes provides an approach for modeling complex decision-making processes by Air Force managers and for eliciting and quantifying their judgments about the value of the outcomes.
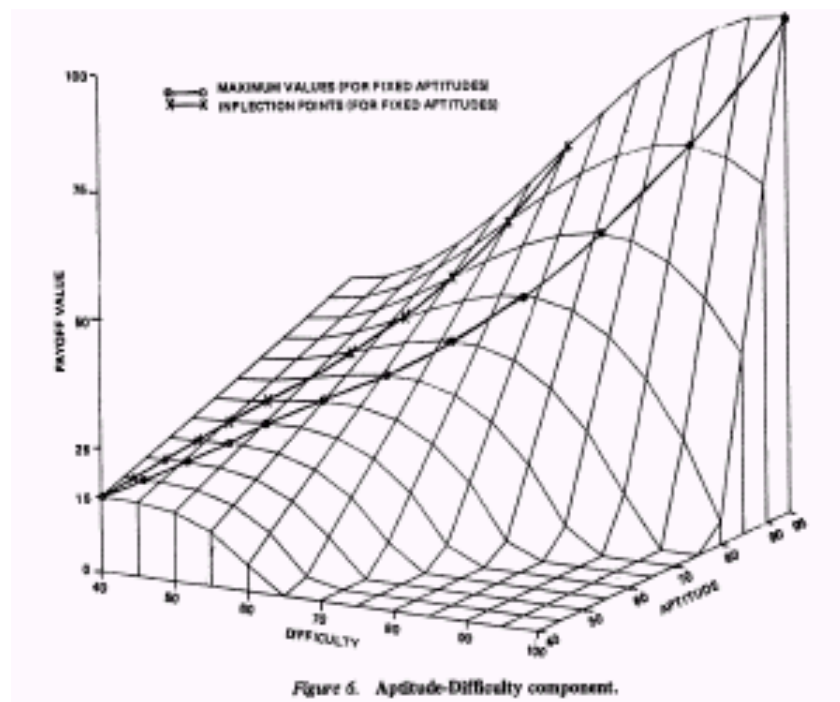
Figure 6. Aptitude-Difficulty component.

Ward, J.H., Jr. (1977). *Creating mathematical models of judgment processes: From policy-capturing to policy-specifying* (AFHRL-TR-77-47). Brooks AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory. **RM-06**

Ward's Clustering Method

Ward's clustering method was developed by Dr. Joe H. Ward Jr. in the 1960s at the Air Force Personnel Research Laboratory. Its purpose was to cluster large numbers of people, objects or symbols into smaller numbers of groups, each group having members that were as much alike (or different) as possible. Ward's clustering method is a general form of hierarchical grouping and is often associated with the use of squared Euclidian distances, the D^2 statistic, as the measure to be optimized in the grouping solution. Data are organized in a symmetric distance matrix containing information about the distances between each x, y pair. The matrix can contain similarity or dissimilarity measures. Different objective functions (minimize, maximize, average) are defined depending on the investigator's research interest.

The original application was to military personnel problems but it was soon adopted by the private sector for grouping and classification problems. The nature of data input makes this clustering method useful in diverse disciplines. Ward's clustering method appeared in one of the earliest statistical analysis software packages for academicians and is available today in the Statistical Analysis System (SAS), Statistical Package for the Social Sciences (SPSS), and other widely used analysis programs. Textbooks on cluster analysis for mathematicians and statisticians cover the methodology.

In the Air Force the method was used in analyses of occupational survey data to describe job types within a specialty. Job incumbents were grouped on similarity of their jobs based on overlap in percent of time spent performing individual tasks. An ongoing study is using the clustering technique to address the feasibility of changes in enlisted specialty structure through mergers based on similarity of skill ratings.

Outside the military the clustering method has been used to establish taxonomies for plants and animals with respect to genetic background and to organize and catalog library documents based on similarity of subject domain to facilitate storage and retrieval of materials. Researchers in the United Kingdom applied Ward's clustering method to 2001 census data to classify the population on demographic structure, household composition, housing, socio-economic character, and employment and industry sectors. The method has also been used in gerontology, chemical, biomedical and dentistry research and studies of hospital governance and international management classifications.

Treat, B.R. (2007). *Hierarchical grouping in Air Force personnel analysis.* San Antonio, TX: Operational Technologies Corporation. **RM-02**

Ward, J.H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, No. 301, 236-244. **RM-06**

Ward, J.H., Jr., Treat, B.R., & Albert, W.G. (1985). *General applications of hierarchical grouping using the HIER-GRP computer program* (AFHRL-TP-84-42). Brooks AFB, TX: Air Force Human Resources Laboratory. **RM-08**

Test Bias Analyses

Beginning with the historic 1954 Supreme Court desegregation order and the passage of the Civil Rights Act of 1964, test psychologists began to address fair test use for minorities. During the next 20 to 30 years, accepted methodologies and definitions emerged. The Air Force contributed to the research stream. Notably, an AFHRL study first used the methodology which became the recommended statistical approach for evaluating test bias.

As a result of the Civil Rights Act of 1964, a federal agency, the Equal Employment Opportunity Commission (EEOC), was established to oversee employers' selection and placement procedures and insure they complied with antidiscrimination laws. In the 1970s the EEOC developed the *Uniform Guidelines on Employee Selection Procedures* which serve as standards for compliance. An important concept defined in the *Uniform Guidelines* is adverse impact. Adverse impact occurs when a personnel decision leads to members of a protected group being excluded from hiring in disproportionate numbers compared to a majority group. The four-fifths rule was established and states that a hiring procedure has adverse impact when the selection rate for any protected group is less than 4/5, or 80 percent, of the group with the highest hiring rate. In an important legal decision, the Supreme Court ruled that the burden of proof on whether an employment selection test is fair rests with the employer. Employers must show that their selection tests and employee selection methods are valid indicators of future job-related performance. The 4/5 rule is used as a threshold for holding the employer accountable for demonstrating that their selection procedures are not biased against any protected group. Otherwise the burden is on the plaintiff.

In the past, the General Accounting Office, Inspector General and Congress have raised issues covered by the *Uniform Guidelines*. In response, the Air Force has made it a practice of determining adverse impact rates, conducting validation studies, and completing test bias analyses. The issues apply to the ASVAB and AFOQT, as well as to tests used for promotion selection decisions.

Cleary's (1968) psychometric model is the most widely used model in the evaluation of test bias. Cleary's definition asserted: "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup." This definition is currently accepted under the *Uniform Guidelines*. Although the method was attributed to Cleary by name, the AFHRL had applied the definition in an Air Force test bias study published 15 years earlier.

The flow chart below serves as an organizational framework for the series of analyses used to address test bias. Following the proper computational procedures is important to insure that **common misunderstandings about test bias** are avoided. For example, bias in testing cannot be assumed as a result of mean differences in test scores alone. Another common error is to inspect the zero-order validity coefficients (*r*) computed within each

comparison group (for example, males and females) and, if they differ, to conclude the test is biased. This practice of comparing validity coefficients is fraught with difficulties, due to the often misleading properties of computed within-group correlations.

```
                        ┌─────────────────────┐
                        │  Test for Equality of│
                        │        SEEs          │
                        └─────────────────────┘
              NS.                              Sig.
        ┌─────────────────────┐      ┌─────────────────────┐
        │  Test for Slope Bias │      │  Conduct analysis   │
        │    Model 1 vs. 2     │      │  using expectancy   │
        └─────────────────────┘      │       tables        │
     Sig.              NS.           └─────────────────────┘
 ┌──────────────────┐   ┌─────────────────────┐
 │  Slope bias      │   │ Test for Intercept Bias│
 │   detected       │   │     Model 2 vs. 3     │
 └──────────────────┘   └─────────────────────┘
                     Sig.               NS.
          ┌──────────────────┐   ┌──────────────────┐
          │ Intercept bias   │   │  No bias detected │
          │   detected       │   │                   │
          └──────────────────┘   └──────────────────┘
```
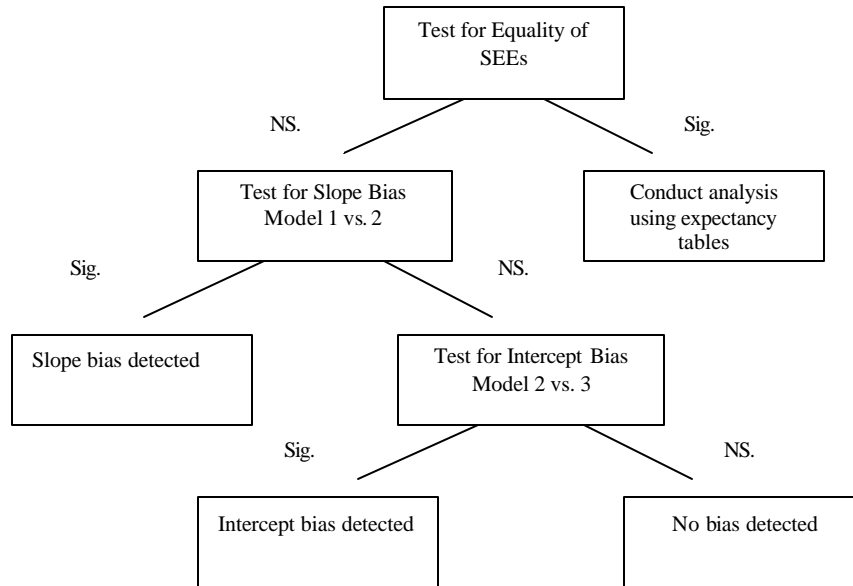
Briefly, computational procedures begin with an overall test for equality of standard errors of estimate (SEE). The comparison of SEEs addresses the natural language question: Is the test measure equally valid for both subgroups? Are the errors of prediction comparable enough within the male and female subgroups, for example, to proceed with parametric comparisons for the Cleary model? Standard errors are immune to some of the more troublesome aspects of validity coefficients and more directly address the differential validity issue.

If the ratio of SEEs shows a statistical difference (Sig.), non-parametric analysis should be conducted. One non-parametric analysis option addresses the question: Is the probability of success at fixed points on the selector the same for both the minority and majority groups? To implement the option, a point on the criterion must be established above which a person is considered successful and below which they would be not successful. Sometimes the criterion has a natural breakpoint (pass/fail), and sometimes the point is identified by management consensus. Expectancy tables are constructed showing the probabilities for success in the two comparison groups at comparable levels on the test measure. The degree of over and underprediction is visually inspected or further significance tests made using the chi-square statistic.

If the test of SEEs is not significant (NS), then parametric tests for the Cleary procedure are followed. The criterion may be defined as a function of generalized linear models. The first or starting model is assumed to be a "true" characterization of the relationships among the expected values in the population. The models subsequently defined represent restricted versions of the starting model and are formed by hypothesizing certain

relationships among the expected values. Comparisons of the degree to which each of the models "fit" the obtained data serve as a test of the null hypotheses. Model comparisons are made sequentially. The first comparison, Model 1 versus 2, tests for slope bias. If significant differences are obtained, then slope bias is detected. If not significant, then the second comparison (Model 2 versus 3) is tested. If the second comparison yields a significant result, then intercept bias is detected. If not significant, then the finding is that no test bias is detected.

The procedures for a test bias study are somewhat complicated, but inspection of mean differences on the test or comparing validity coefficients that might be subject to sampling variations will not suffice.

Alley, W.E. (2006, November). *Conducting a test bias study*. San Antonio, TX: Operational Technologies Corporation. **RM-09**

Linear Models

Linear models are mathematical tools that take the form of equations solvable by multiple regression analyses with the outcome measure (measure of interest or dependent variable) on the left side of the equation, and the predictor information (independent variables) on the right side. Relationships among expected values can be modeled and tested so that a) estimates of expected values can be obtained from the independent variables and b) specific hypothesized relationships among the expected values can be tested using the F-ratio and associated probability values. The Air Force Human Resources Laboratory (AFHRL) made extensive use of the general linear models approach to statistical analyses. Linear models provide a flexible approach to analyzing Air Force personnel research questions, many of which are not addressable using traditional and often inflexible experimental designs like those commonly known as t-test, analysis of variance, and analysis of covariance. The latter statistical methods are simply special cases of the general approach called linear models or multiple linear regression. The use of linear models allow Air Force researchers to formulate and solve often complex research questions without concerns about "matching experimental subjects," "equating cell frequencies" and other rigid requirements of "standard" designs.

Linear models provide a direct and powerful approach to the effective formulation and resolution of a wide variety of research problems. The assumptions underlying the regression approach are less restrictive. Predictor variables, for example, are not assumed to come from multivariate normal distributions. Hence, one advantage of this approach is that it is admirably suited to problems in which predictive information is in the form of binary-coded (1 or 0) data such as gender or high school graduate or non-graduate. The approach also accommodates the specification of a large number of continuous and categorical predictor variables with polynomial forms for large samples, situations that often arise in Air Force research using historical personnel data bases. The approach provides a technique for researchers to pose their research questions in natural language form, to then express them correctly as a general linear mathematical model, and finally to test hypotheses about competing models through comparisons of alternate models (full and restricted) with the F-statistic.

The applications of the general linear models approach in AFHRL research are too numerous to list. Suffice it to say that nearly all large studies of enlisted and officer selection and classification problems used the technique. Although most graduate courses on research methodologies and statistics still focus on traditional experimental designs, the general linear models approach adopted by the Air Force many decades ago is gradually making its way into university and even high school curricula. Some of the progress can be attributed to the large number of researchers trained at AFHRL over the years, many of whom are now filling positions as college professors and advisers to high school science programs.

Bottenberg, R.A., & Ward, J.H., Jr. (1963). *Applied multiple linear regression* (PRL-TDR-63-6). Lackland AFB, TX: Personnel Research Laboratory. **RM-10**

KEYWORD INDEX

A

Job spin process 48
Joint-Service Job Performance Measurement (JPM) Project 62-63

L

Leadership Effectiveness and Assessment Profile (LEAP) 106
Learning Abilities Measurement Program (LAMP) 66-69, 101-102
Learning difficulty 43-44
Learning-skills taxonomy 66
Linear models 134
Literacy 31-32, 40
Logic gates 66
Low aptitude recruits 31-34

M

MAGE (Mechanical, Administrative, General, Electronic) 11, 13, 34, 42-43, 51, 68, 125
Marital status 56, 59-60
Methodology for Identifying Abilities for Job Specialties (Project MIDAS) 105
Minimum qualifying scores 43-44, 53, 89
Misnorming 33, 57
Moral waivers 61
Multiple linear regression 134

N

National Longitudinal Study of Youth Labor Force Behavior 22, 34
Navigator selection 79, 82-84, 87-89, 91-96, 101, 103-104
Navigator-Technical composite utility analysis 95-96
Navy Vocational Interest Inventory 15
New Standards Men 31-32
Norms 22-23, 33
Numerical Operations (NO) Test 14, 16, 17, 19-20, 26, 52

O

Occupational Attitude Inventory (OAI) 75
Occupational learning difficulty 43
Occupational surveys 43, 75-77, 100
Officer ability requirements 105
Officer Biographical Inventory 81-84
Officer Effectiveness Report (OER) 118-120
Officer grade requirements 121-122, 127
Officer grade structure 121-122, 127
Officer Performance Report (OPR) 118-120
Officer screening tests 97

Vocational interests 15, 48, 51, 70

W

Walk-through Performance Test (WTPT) 64-65
Ward's clustering method 130
Weighted Airman Promotion System (WAPS) 71-74, 125-126
Weighted Professional Officer Course Selection System (WPSS) 115, 127
Women in the Air Force 10-11, 55-58, 60, 84, 94, 131-133
Work sample tests 62, 64
Working memory capacity 51, 66-68
World War II mobilization norms 22

AUTHOR INDEX

A

Albert, W.G.	130
Alderton, D.L.	26
Alley, W.E.	14, 29, 50, 52, 53, 61, 70, 88, 94, 111, 133
Androlewicz, T.R.	18
Appel, V.A.	106

B

Barrick, M.R.	110
Berger, F.R.	83
Berger, R.M.	83
Bloxom, B.M.	26
Bock D.R.	35
Bottenberg, R.A.	120, 125, 134

C

Campbell, J.A.	27
Campbell, J.P.	54
Carretta, T.R.	52, 98, 99, 100, 102, 104, 116
Chaiken, S.R.	69
Chiorini, J.R.	104
Christal, R.E.	110, 122, 125, 127
Cole, R.W.	106
Creager, J.A.	80, 81, 83, 85
Curran, L.T.	18, 28, 104
Curton, E.D.	120

D

Darby, M.M.	53
Diaz, T.D.	52
DiFazio, A.S.	27
Dittmar, M.J.	105
Divgi, D.R.	30
Driskill, W.E.	105

E

Earles, J.A.	21, 23, 30, 32, 35, 69, 83, 85, 105, 116
Eitelberg, M.J.	35

L

M

N

P

R

SUPPLEMENTAL READING

Alley, W.E. (2002). *Development of experimental measures for the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-06**

Alley, W.E. (2006). *Conducting a test bias study*. San Antonio, TX: Operational Technologies Corporation. **RM-09**

Alley, W.E., & Darby, M.M. (1995). Estimating the benefits of personnel selection and classification: An extension of the Brodgen table. *Educational and Psychological Measurement, 55(6)*, 938-958. **E-30**

Alley, W.E., & Matthews, M.D. (1982). The Vocational Interest Career Examination: A description of the instrument and possible applications. *The Journal of Psychology, 112*, 169-193. **E-37**

Alley, W.E., & Teachout, M.S. (1995). *Differential assignment potential in the ASVAB: A simulation of job performance gains* (AL/HR-TP-1995-0006). Brooks AFB, TX: Human Resources Directorate, Armstrong Laboratory. **E-27**

Barrick, M.R., & Mount, M.K. (2005). Yes, personality matters: Moving on to more important matters. *Human Performance*, *18*(4), 359-372. **O-30**

Berger, F.R., Gupta, W.B., Berger, R.M., & Skinner, J. (1990, April). *Air Force Officer Qualifying Test (AFOQT) Form P: Test Manual* (AFHRL-TR-89-56, AD-A221 004). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-03**

Bock, D.R., & Moore, E.G.J. (1984). *The Profile of American Youth: Demographic influences on ASVAB test performance* (DTIC Report No. AD-A125 830). Washington, D.C.: Office of the Assistant Secretary for Defense (Manpower, Installations, and Logistics). **E-48**

Bottenberg, R.A. (1978). *Relationships among factors in new officer effectiveness report system* (AFHRL-TR-78-40). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-37**

Bottenberg, R.A., & Christal, R.E. (1961). *An iterative technique for clustering criteria which retains optimum predictive efficiency* (WADD-TN-61-30). Lackland AFB, TX: Personnel Laboratory. **RM-01**

Bottenberg, R.A., & Ward, J.H., Jr. (1963). *Applied multiple linear regression* (PRL-TDR-63-6). Lackland AFB, TX: Personnel Research Laboratory. **RM-10**

Campbell, J.P., Russell, T.L., & Knapp, D.J. (1994). *Roadmap: An agenda for joint-service classification research* (AL/HR-TP-1994-0003). Brooks AFB, TX: Human Resources Directorate, Armstrong Laboratory. **E-31**

Carretta, T.R. (1987). *Basic Attributes Tests (BAT) system: Development of an automated test battery for pilot selection* (AFHRL-TR-87-9, ADA185649). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **O-23**

Carretta, T.R. (2005). *Development and validation of the Test of Basic Aviation Skills* (AFRL-HE-WPTR-2005-0172, ADA442563). Wright-Patterson AFB, OH: Human Effectiveness Directorate, Air Force Research Laboratory. **O-18**

Carretta, T.R., & Ree, M.J. (1993) *Pilot candidate selection method (PCSM): What makes it work?* (AL-TP-1992-0063). Brooks AFB, TX: Manpower and Personnel Research Division, Human Resources Directorate, Armstrong Laboratory. **O-17**

Carretta, T.R., Perry, D.C., Jr., & Ree, M.J. (1997). Prediction of situational awareness in F-15 pilots. *The International Journal of Aviation Psychology, 6(1)*, 21-41. **O-43**

Christal, R. E. (1967). *Selecting a harem and other applications of the policy-capturing model* (PRL-TR-67-1). Lackland AFB, TX: Personnel Research Laboratory. **RM-04**

Christal, R.E. (1975). *Systematic methods for establishing officer grade requirements based upon job demands* (AFHRL-TR-75-36, AD-A015 756). Lackland AFB, TX: Air Force Human Resources Laboratory. **O-41**

Christal, R.E. (1992). Author's note on "Recurrent personality factors based on trait ratings." *Journal of Personality, 60(2)*, 221-224. **O-28**

Curran, L.T., & Palmer, P. (1990). *Armed Services Vocational Aptitude Battery (ASVAB): Item, overlength and operational length development of Forms 18 and 19.* (AFHRL-TP-89-74). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **E-08**

Defense Manpower Data Center (2005). ASVAB Career Exploration Program Counselor Manual. North Chicago, Il: HQ USMEPCOM. **E-44**

Diaz, T., Ingerick, M., & Lightfoot, M.A. (2004). *Replication of Zeidner, Johnson, and colleagues' method for estimating Army aptitude area (AA) composites* (DTIC Report No. ADA 426 299). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences. **E-28**

Directorate of Accession Policy, Office of the Under Secretary (Personnel and Readiness) (2006). *Population representation in the military services, Fiscal Year 2004.* Washington, D.C.: Author. **E-33**

Dittmar, M.J., Weissmuller, J.J., Driskill, W.E., Hand, D.K., & Earles, J.A. (1994). *Methodology for identifying abilities for job specialties (MIDAS)* (AL/HR-TP-1994-

0008, AD-A277 919). Brooks AFB, TX: Human Resources Directorate, Armstrong Laboratory. **O-24**

Earles, J.A., & Ree, M.J. (1998). *Development and evaluation of alternative Air Force Administrative composites from the ASVAB.* Unpublished manuscript. Brooks AFB, TX: Human Effectiveness Directorate, Air Force Research Laboratory. **E-45**

Eitelberg, M.J., Laurence, J.H., Waters, B.K., & Perelman, L.S. (1984). *Screening for Service: Aptitude and education criteria for military entry.* Washington, D.C.: Office of Assistant Secretary of Defense (Manpower, Installations and Logistics). Electronic copy not available.

Fast, J.C., & Looper, L.T. (1988). *Multiattribute decision modeling techniques: A comparative analysis* (AFHRL-TR-88-3). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **RM-05**

Finsteun, K., & Alley, W.E. (1983). *Occupational correlates of first term enlisted tenure* (AFHRL-TR-82-36). Brooks AFB, TX: Air Force Human Resources Laboratory. **E-34**

Finstuen, K., Matthews, G.N. & Pope, W.H. (1980). *Management engineering team applications of officer grade requirements method* (AFHRL-TR-80-32, AD-A093 508). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-42**

Glomb, T.M., & Earles, J.A. (1997). *Air Force Officer Qualifying Test (AFOQT) Forms Q: Development, Preliminary Equating and Operational Equating* (AL/HR-TP-1996-0036). Brooks AFB, TX: Armstrong Laboratory. **O-04**

Gould, R.B. (1978). *Air Force Officer Qualifying Test Form N: Development and standardization* (AFHRL-TR-78-43, AD-A059 746). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-05**

Gould, R.B. (1976). *Review of an Air Force job satisfaction research project: Status report through September 1976* (AFHRL-TR-76-75). Lackland AFB, TX: Air Force Human Resources Laboratory**. E-16**

Grobman, J.H., Alley, W.E., & Pettit, R.S. (1995). *The optimality of sequential personnel assignments using a Decision Index* (AL/HR-TP-1995-0026). Brooks AFB, TX: Human Resources Directorate, Armstrong Laboratory. **E-38**

Guinn, N., Tupes, E.C., & Alley, W.E. (1970). *Cultural subgroup differences in the relationships between Air Force aptitude composites and training criteria* (AFHRL-TR-70-35). Brooks AFB, TX: Air Force Human Resources Laboratory. **E-40**

Ingerick, M. (2005). *Identifying leader talent: Alternative predictors for U.S. Air Force junior officer selection and assessment* (DFR-05-47). Alexandria, VA: Human Resources Research Organization. **O-31**

Jensen, H. E., Massey, I.H., & Valentine, L.D. Jr. (1976). *Armed Services Vocational Aptitude Battery Development (ASVAB Forms 5, 6, and 7)*. (AFHRL-TR-76-87, AD-AD-AO37 522). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **E-15**

Johnson, C.A., Meehan, J., & Wilkinson, R.E. (1976). *Officer effectiveness report development – 1971 through 1972* (AFHRL-TR-76-61). Lackland AFB, TX: Air Force Human Resources Laboratory. **O-38**

Kyllonen, P.C. (2007). *The Learning Abilities Measurement Program (LAMP) 1982-1999*. San Antonio, TX: Operational Technologies Corporation. **E-03**

Laurence, J.H., Naughton, J., & Harris, D.A. (1996). *Attrition revisited: Identifying the problem and its solutions* (ARI Research Note 96-20). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences. **E-35**

Laurence, J.H., Ramsberger, P.F., & Gribben, M.A. (1989). *Effects of military experience on the post-service lives of low-aptitude recruits: Project 100,000 and the ASVAB misnorming* (HumRRO FR-PRD-89-29, Educational Resources Information Center (ERIC) Document Number ED 366 751). Alexandria, VA: Human Resources Research Organization. **E-12**

Maier, M.H., & Fuchs, E.F. (1973). *Effectiveness of selection and classification testing* (ARI-RR-1179, AD0768168). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences. **E-49**

Martin, C.J. & Welsh, J.R. (1999). Comparison of 1980 and 1997 ASVAB norming procedures. Monterey, CA: *Proceedings of the 41st Annual Conference of the International Military Testing Association*, p. 329. **E-05**

McCloy, R.A., Campbell, J.A., Knapp, D.J., Strickland, W.J., & DiFazio, A. S. (2006). *A framework for conducting validation research with the Armed Services Vocational Aptitude Battery (ASVAB)*. Alexandria, VA: HumRRO. **E-10**

McCrae, R.P. (1992). Editor's introduction to Tupes and Christal. *Journal of Personality, 60*(2), 217-219. **O-28**

McCrae, R.P., & John, O.P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175-215. **O-28**

Miller, R.E. (1969). *Interpretation and utilization of scores on the Air Force Officer Qualifying Test* (AFHRL-TR-69-103, AD 691 001). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **O-10**

Nataupsky, M., Curton, E.D., Waller, E., & Ratliff, F.R. (1977). *Report on the 1975 officers' OER opinion survey* (AFHRL-TR-77-37). Brooks AFB, TX: Air Force Human Resources Laboratory. **O-39**

Palmer, P., Curran, L., & Haywood, C.S. (1990). *Armed Services Vocational Aptitude Battery (ASVAB) Forms 20, 21, and 22: Item development.* (AFHRL-TP-89-77). Manpower and Personnel Research Division, Air Force Human Resources Laboratory. **E-41**

Piña, M., Jr. (2007). *Pre and post enlistment person job match (PJM).* San Antonio, TX: Operational Technologies Corporation. **E-26**

Prestwood, J.S., & Vale, C.D. (1985). *Armed Services Vocational Aptitude Battery: Development of Forms 11, 12, and 13*. (AFHRL-TR-85-16(I)). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **E-42**

Ratliff, F.R., & Earles, J.A. (1976). *Research on the management, training, and utilization of low-aptitude personnel* (AFHRL-TR-76-69). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **E-13**

Ratliff, F.R., Chiorini, J.R., Curran, C.R., & Shore, C.W. (1970, November). *Evaluating combat crew training performance using criteria of minimum performance standards* (AFHRL-TR-70-50). Lackland AFB, Texas: Personnel Research Division, Air Force Human Resources Laboratory. **O-45**

Ratliff, F.R., Shore, C.W., Chiorini, J.R., & Curran, C.R. (1969, July). *Inflight performance differences of pilot and navigator F-4 second-seat crew-members: A limited Southeast Asia combat evaluation* (AFHRL-TR-69-104, AD-705 140). Lackland AFB, Texas: Personnel Research Division, Air Force Human Resources Laboratory. **O-46**

Ree, M. J., & Earles, J.A. (1998). *Development and evaluation of alternative Air Force Mechanical composites from the ASVAB.* Unpublished manuscript. Brooks AFB, TX: Human Effectiveness Directorate, Air Force Research Laboratory. **E-46**

Ree, M.J. (2003a). *Test of Basic Aviation Skills (TBAS): Incremental validity beyond the Air Force Officer Qualifying Test (AFOQT) Pilot composite for predicting pilot criteria.* San Antonio, TX: Operational Technologies Corporation. **O-21**

Ree, M.J. (2003b). *Test of Basic Aviation Skills (TBAS): Scoring the tests and compliance of the tests with the standards of the American Psychological Association.* San Antonio, TX: Operational Technologies Corporation. **O-22**

Ree, M.J. (2004a). *Making scores equivalent for Test of Basic Aviation Skills (TBAS) and Basic Attributes Test (BAT)*. San Antonio, TX: Operational Technologies Corporation. **O-19**

Ree, M.J. (2004b). *Reliability of the Test of Basic Aviation Skills* TBAS. San Antonio, TX: Operational Technologies Corporation. **O-20**

Ree, M.J., & Carretta, T.R. (2002). g2K. *Human Performance, 15*(1/2), 3-23. **E-29**

Ree, M.J., Carretta, T.R., & Earles, J.A. (2003). Salvaging construct equivalence through equating. *Personality and Individual Differences*, 35 (6), 1293-1305. **O-35**

Ree, M.J., Mathews, J.J., Mullins, C.J., & Massey, R.H. (1981). *Calibration of Armed Services Vocational Aptitude Battery Forms 8, 9, and 10.* (AFHRL-TR-81-49). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **E-06**

Ree, M.J., Welsh, J.R., Wegner, T.G., & Earles, J.A. (1985). *Armed Services Vocational Aptitude Battery: Equating and implementation of Forms 11, 12, and 13 in the 1980 youth population metric* (AFHRL-TP-85-21). Brooks AFB, TX: Manpower and Research Division, Air Force Human Resources Laboratory. **E-07**

Rogers, D.L., Roach, B.W., & Short, L.W. (1986). *Mental ability testing in the selection of Air Force Officers: A brief historical overview*. (AFHRL-TP-86-23). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **O-01**

Romaglia, D.L., & Skinner, J. (1990). Validity of grade point average: Does the college make a difference. *Proceedings of the 32nd Annual Conference of the Military Testing Association*, p. 345-350. Orange Beach, AL. **O-33**

Sands, W.A., Gade, P.A., & Knapp, D.J. (1997). The Computerized Adaptive Screening Test. In Sands, W.A., Waters, B.K., & McBride, J.R. (Eds.) *Computerized adaptive testing: From inquiry to operation*. Washington, D.C.: American Psychological Association. (Also published as HumRRO FR-EADD-96-26) **E-09**

Sands, W.A., Waters, B.K., & McBride, J.R. (Eds.) *Computerized adaptive testing: From inquiry to operation*. Washington, D.C.: American Psychological Association. (Also published as HumRRO FR-EADD-96-26) **E-09**

Sawin, L., Earles, J., Goff, G.N., & Chaiken, S.R. (2001). *Advanced personnel testing project: Final report* (AFRL-HE-AZ-TP-2001-0004). Brooks AFB, TX: Warfighter Training Research Division, Air Force Research Laboratory. **E-36**

Sellman, W.S. (2004). *Predicting readiness for military service: How enlistment standards are established.* Prepared for The National Assessment Governing Board. **E-47**

Shermis, M.D., Falkenberg, B., Appel, V.A., & Cole, R.W. (1996). Construction of a faking detector scale for a biodata survey instrument. *Military Psychology, 8* (2), 83-94. **O-26**

Shore, C.W., & Gould, R.B. (2004). *Revalidation of WAPS and SNCOPP.* San Antonio, TX: Operational Technologies Corporation. **E-25**

Shore, C.W., & Marion, R. (1972). *Suitability of using common selection test standards for Negro and White airmen* (AFHRL-TR-72-53). Lackland AFB, Texas: Personnel Research Division, Air Force Human Resources Laboratory. **E-50**

Shore, C.W., & Gould, R.B. (2002). *Reduction of Air Force Officer Qualifying Test (AFOQT) administration time.* San Antonio, TX: Operational Technologies Corporation. **O-08**

Shore, C.W., & Gould, R.B. (2003). *Developing pilot and navigator/technical composites for the Air Force Officer Qualifying Test (AFOQT) Form S* San Antonio, TX: Operational Technologies Corporation. **O-07**

Shore, C.W., Curran, C.R., Ratliff, F.R., & Chiorini, J.R. (1970, April). *Proficiency differences of pilot and navigator F-4 second-seat crew members: A Southeast Asia evaluation* (AFHRL-TR-70-9, AD-709 728). Lackland AFB, Texas: Personnel Research Division, Air Force Human Resources Laboratory. **O-44**

Skinner, J. (2006). *Air Force Reserve Officer Training Corps (AFROTC) studies*. San Antonio, TX: Operational Technologies Corporation. **O-34**

Skinner, J. (2007). *Substituting commercial tests for the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-32**

Skinner, J. (2006). *AFOQT retesting*. San Antonio, TX: Operational Technologies Corporation. **O-13**

Skinner, J. (2006). *Criterion-related validity of the operational AFOQT composites*. San Antonio, TX: Operational Technologies Corporation. **O-12**

Skinner, J. (2006). *Officer screening tests.* San Antonio, TX: Operational Technologies Corporation. **O-16**

Skinner, J. (2006). *Strength Aptitude Test*. San Antonio, TX: Operational Technologies Corporation. **E-22**

Skinner, J. (2006). *Test bias studies of the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-14**

Skinner, J. (2007). A*ir Force Reading Abilities Test (AFRAT) and related topics.* San Antonio, TX: Operational Technologies Corporation. **E-17**

Skinner, J. (2007). *Educational level norms.* San Antonio, TX: Operational Technologies Corporation. **O-11**

Skinner, J. (2007). *Enlistment Screening Test (EST) and Computerized Adaptive Screening Test (CAST).* San Antonio, TX: Operational Technologies Corporation. **E-19**

Skinner, J. (2007). *Joint-service classification research roadmap.* San Antonio, TX: Operational Technologies Corporation. **E-32**

Skinner, J. (2007). *New content areas for the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-25**

Skinner, J. (2007). *Occupational learning difficulty for establishing aptitude requirements.* San Antonio, TX: Operational Technologies Corporation. **E-23**

Skinner, J. (2007). *Substituting commercial tests for the AFOQT*. San Antonio, TX: Operational Technologies Corporation. **O-32**

Skinner, J. (2007). *Utility analysis of the AFOQT Pilot and Navigator Composites*. San Antonio, TX: Operational Technologies Corporation. **O-15**

Skinner, J. (2007). *Electronic Data Processing Test (EDPT)*. San Antonio, TX: Operational Technologies Corporation. **E-20**

Sturiale, G. (1968). *The officer effectiveness report as a performance measure: A research review* (AFHRL-TR-68-113). Lackland AFB, TX: Air Force Human Resources Laboratory. **O-36**

Syllogistics Inc. and The Hay Group (1987*). Final report: Air Force officer evaluation system project* (DTIC Report No. AD-A223 834). Springfield, VA. **O-40**

Teachout, M.S. (2007). *The Joint-Service Job Performance Measurement/Enlistment Standards Project and the Air Force Job Performance Measurement Project: A summary of key results.* San Antonio, TX: Operational Technologies Corporation. **E-21**

Thompson, N. (2007). *Enlisted selection and classification tests: Precursors of the ASVAB.* San Antonio, TX: Operational Technologies Corporation. **E-01**

Thompson, N. (2007). *Tracking trends in the enlisted force*. San Antonio, TX: Operational Technologies Corporation. **E-14**

Treat, B.R. (2007). *Hierarchical grouping in Air Force personnel analysis*. San Antonio, TX: Operational Technologies Corporation. **RM-02**

Tupes, E.C., & Christal, R.E. (1961) *Recurrent personality factors based on trait ratings*. (ASD-TR-61-97). Lackland AFB, TX: Personnel Laboratory, Aeronautical Systems Division. **O-29**

Tupes, E.C., & Christal, R.E. (1992). Recurrent personality factors based on trait ratings. *Journal of Personality, 60*(2), 225-251. **O-28**

Tuttle, T.C., Gould, R.B., & Hazel, J.T. (1976). *Dimensions of job satisfaction: Initial development of the Air Force Occupational Attitude Inventory* (AFHRL-TR-75-1). Lackland AFB, TX: Air Force Human Resources Laboratory. **E-18**

Valentine, L.D., Jr., & Creager, J.A. (1961). *Officer selection and classification tests: Their development and use*. (ASD-TN-61-145). Lackland AFB, TX: Personnel Laboratory, Aeronautical Systems Division. **O-02**

Vitola, B.M., & Alley, W.E. (1968). *Development and standardization of Air Force Composites for the Armed Services Vocational Aptitude Battery*. (AFHRL-TR-68-10). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **E-04**

Ward, J.H., Jr. (1959). *Use of a Decision Index in assigning Air Force personnel* (WADC-TN-59-38). Lackland AFB, TX: Personnel Laboratory. **E-39**

Ward, J.H., Jr. (1961). *Hierarchical grouping to maximize payoff* (WADD-TN-61-29, AD-261 750). Lackland AFB, TX: Personnel Laboratory. **RM-03**

Ward, J.H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, No. 301, 236-244. **RM-06**

Ward, J.H., Jr. (1977). *Creating mathematical models of judgment processes: From policy-capturing to policy-specifying* (AFHRL-TR-77-47). Brooks AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory. **RM-06**

Ward, J.H., Jr., Treat, B.R., & Albert, W.G. (1985). *General applications of hierarchical grouping using the HIER-GRP computer program* (AFHRL-TP-84-42). Brooks AFB, TX: Air Force Human Resources Laboratory. **RM-08**

Weeks, J. (1984). *Occupational learning difficulty: A standard for determining the order of aptitude requirement minimums* (AFHRL-SR-84-26). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory. **E-24**

Weeks, J.L. (2000). *USAF pilot selection* (AFRL-HE-AZ-TP-2000-0004). Mesa, AZ: Warfighter Training Research Division, Air Force Research Laboratory. **O-27**

Weeks, J.L., Mullins, C.J., & Vitola, B.M. (1975). *Airman Classification Batteries from 1948 – 1975: A review and evaluation* (AFHRL-TR-75-78). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. **E-02**

Weissmuller, J.J., Schwartz, K.L., Kenney, S.D., Shore, CW., & Gould, R.B. (2004). Recent developments in USAF officer testing and selection. *Proceedings of 46th Annual Conference of the International Military Testing Association and NATO's Research Task Group on Recruiting and Retention of Military Personnel*, p. 268-279. Brussels, Belgium. **O-09**

Welsh, J.R., Androlewicz, T.R., & Curran, L.T. (1990). *Armed Forces Vocational Aptitude Battery (ASVAB): Analysis of differential item functioning on Forms 15, 16, and 17.* (AFHRL-TP-90-62). Brooks AFB, TX: Manpower and Personnel Research Division, Air Force Human Resources Laboratory. **E-43**

Welsh, J.R., Jr., Kucinkas, S.K., & Curran, L.T. (1990). *Armed Services Vocational Aptitude Battery (ASVAB): Integrative review of validity studies* (AFHRL-TR-90-22). Brooks AFB, TX: Air Force Human Resources Laboratory. **E-11**

Wise, L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L., & Divgi, D.R. (1992). *Sensitivity and fairness of the Armed Services Vocational Aptitude Battery (ASVAB) technical composites.* Seaside, CA: Defense Manpower Data Center. **E-51**

Wolfe, J.H., Alderton, D.L., Larson, G.E., Bloxom, B.M., & Wise, L.L. (1997). Expanding the content of CAT-ASVAB: New tests and their validity. In Sands, W.A., Waters, B.K., & McBride, J.R. (Eds.) *Computerized adaptive testing: From inquiry to operation.* Washington, D.C.: American Psychological Association. (Also published as HumRRO Report No. FR-EADD-96-26.) **E-09**